Aalto University
School of Science
Degree Programme of Engineering Physics and Mathematics

Ilkka Mansikkamäki

Histogram based signatures for detecting warranty fraud

Masters thesis submitted in partial fulfillment of the requirements for the degree programme of Master of Science in Technology in the Degree Programme in Engineering Physics and Mathematics.

Espoo, August 13, 2012

Supervisor: Professor Ahti Salo
Instructor: Dr.Tech Terhi Pere

AALTO UNIVERSITY
School of Science

ABSTRACT OF
MASTER'S THESIS

Author: Ilkka Mansikkamäki

Title:
Histogram based signatures for detecting warranty fraud

Title in Finnish:
Histogrammeihin perustuvat allekirjoitukset takuupetosten havaitsemisessa

Degree Programme:
Degree Programme in Engineering Physics and Mathematics

Major: Systems and Operations Research

Minor: Industrial Management

Chair (Code): Mat-2

Supervisor: Professor Ahti Salo **Instructor:** Doc.Tech. Terhi Pere

Companies in all industries lose billions of dollars due to fraud every year. Fraud occurs in various ways, one being warranty fraud by repairing partners who invent their repair data. The company that has outsourced its repair activities must have proper tools for detecting fraud. However, it is often impossible to determine single repair data points as fraud, making it profitable for the company to determine the reliability of the repair vendor by investigating the overall performance. This thesis focuses on estimating the reliability of the vendor by comparing the performance of each vendor against others.

This thesis introduces a histogram based profiling method that can be used for vendor comparison as a profile over a period of time or by updating the profile constantly and recording the changes in performance. Profiles, called histogram signatures, are applied to clustering and local outlier methods. Histogram signatures are also compared to identify the changes in the profiles of each vendor's peers. Histograms are compared with Jensen-Shannon divergence difference.

The presented histogram method is tested with real repair data from an electronics company. Fraudulent repair data is simulated to represent different fraud types. The results show that single and momentary changes in the profile are not detected but the method is able to detect well big changes in repairing activity.

Date: August 13, 2012 **Language:** English **Pages:** 65

Keywords: warranty, fraud, histogram, profiling, signatures, clustering, behavior model, change detection

AALTO-YLIOPISTO
Perustieteiden korkeakoulu

DIPLOMITYÖN
TIIVISTELMÄ

Tekijä: Ilkka Mansikkamäki

Työn nimi:
Histogrammeihin perustuvat allekirjoitukset takuupetosten havaitsemisessa

Työn nimi englanniksi:
Histogram based signatures for detecting warranty fraud

Tutkinto-ohjelma:
Teknillisen fysiikan ja matematiikan tutkinto-ohjelma

Pääaine: Systeemi- ja operaatiotutkimus

Sivuaine: Teollisuustalous

Opetusyksikön koodi: Mat-2

Työn valvoja: Professori Ahti Salo **Työn ohjaaja:** TkT Terhi Pere

Yritykset menettävät petoksen takia miljardeja dollareja vuosittain. Petosta ilmenee monin tavoin, kuten esimerkiksi huoltopartnerien tekaistulla takuunalaisella korjausdatalla. Huoltotoimintansa ulkoistaneella yrityksellä pitääkin olla kattava tietojärjestelmä petoksen havaitsemiseksi. Yksittäisiä huoltotapahtumia harvoin voidaan varmuudella arvioida petokseksi, joten yrityksen on kannattavaa tarkastella partnerin toiminnan luotettavuutta kokonaisuudessaan. Tässä työssä määritetään partnerien epäilyttävyyttä vertaamalla kunkin huoltopartnerin toimintaa toisiin partnereihin.

Työssä esitellään histogrammeihin perustuva profilointimenetelmä, jolla huoltopartnerien toimintaa voidaan verrata sekä pitkän aikavälin että päivitettävänä, toiminnan muuttumista tarkkailevana profiilina. Histogrammimenetelmää sovelletaan klusterointi- ja paikallisten poikkeamien havaitsemismenetelmiin sekä tutkitaan profiilin muutoksia lähimpiin naapureihin verrattuna. Histogrammien vertailuun käytetään Jensen-Shannonin divergenssimittaa.

Esiteltyä histogrammimenetelmää testataan elektroniikkayrityksen huoltodatalla, johon generoidaan erilaisia poikkeuksellista huoltotoimintaa mallintavia datapisteitä. Tuloksista käy ilmi, että menetelmä havaitsee heikosti yksittäisiä poikkeamia mutta hyvin suuria muutoksia huoltopartnereiden toiminnassa.

Päivämäärä: 13.08.2012 **Kieli:** Englanti **Sivumäärä:** 65

Avainsanat: petoksen havaitseminen, histogrammi, profilointi, klusterointi käyttäytymismalli, muutoksen havaitseminen

Acknowledgements

During the writing process of this thesis I have received great support from many directions. I want to thank my employing company that gave me the opportunity to write my thesis on this highly interesting topic. I want to thank Professor Ahti Salo for supervision and valuable comments. I have had a couple of supervisors during my thesis; thank you Susanna Alaja and Terhi Pere for your comments and guidance. I also want to express my gratitude to Tom Browne and Elina Suhonen, who took the time and effort to read my writings and suggest improvements.

Finally, I want to thank my colleagues for supporting me with the tools and data. Thanks to my friends for throwing me tips here and there, being supportive and pushing me to get moving with this thesis. Last but definitely not least, I thank my family for their support and patience during my studies and thesis writing process. It gives strength and motivation to know that the people closest to me believe in me.

Espoo, 13.08.2012

Ilkka Mansikkamäki

Contents

Chapter 1

Introduction

1.1 Background

When a product is malfunctioning, consumers often seek compensation from the manufacturer. Typically the compensation is a properly functioning device, i.e., the device is repaired and returned to the consumer. The compensation rule is stated by the consumer law. When the product is purchased, the manufacturer has an obligation to compensate for products that fail within a fixed time period called the warranty period. If the warranty period has already ended, the monetary responsibility of the repair usually belongs to the consumer. When there is a third party operating between the manufacturer and the consumer, repairing the devices, the process of finding the responsible parties is more difficult than when companies deal directly with the consumer. These repairing vendors act as the consumer interface and, from the manufacturers' point of view, are the ones creating the warranty claims. The problem in having these repair vendors is that it leaves possibilities for dishonest warranty claiming. In the worst case, dishonesty turns to vendor fraud.

Fraud creates major expenses for companies. In 2009, companies worldwide lost about 4.6% of their expenditure, 2.7 trillion dollars in total, due to fraud [20]. Most common fraud types include credit card, health care, telemarketing and online advertising fraud, identity theft, software piracy and numerous other scams [42]. No industry has avoided fraud completely. Where there is a process involving trust, there is a party breaching the mutual confidence to gain some unfair or dishonest advantage, i.e. performing fraud [16].

When a warranty claim is made by a repair partner against the manufacturing company, the claim is valid by default. However, as stated above, the partnership is easily exposed to fraud. Warranty claims sent by the repair vendors are sometimes difficult to validate, because the information required for detailed validation is hard to acquire. Furthermore, the validation of a claim is a complex and time consuming process when the claimant is assumed trustworthy and the repair volume is so great that manual validation is not possible. The difficulty of validation arises when the third party has access to several warranty claim items simultaneously, thus either inventing non-existing claims or reporting higher value claims than necessary, for example. Several other examples of fraud in warranty claims exist and they evolve constantly as the fraud detection methods become more sophisticated.

To protect themselves against fraud, organizations have developed controlling techniques that seek to prevent fraudulent activity before it causes financial impact for the company. The controlling techniques can broadly be divided into technical and non-technical controls. The former are generally operational and predictive mathematical models, the latter focus on detecting suspicious human behavior. Non-technical controls take into account that for each process there is a human operator [21]. Here lies the challenge of fraud detection; how to combine technical and non-technical controls and detect suspicious human behavior with automated technical controls? This thesis seeks to answer this question by studying a chosen mathematical method for behavior identification. In order to understand the purpose of these methods, a comprehensive perspective of fraud is needed.

In addition to the behavioral view of fraud, authors have approached fraud control from other perspectives. One view divides fraud controlling tasks to prevention and detection. Fraud prevention is described as *the measures to stop fraud occurring* and fraud detection as *identifying fraud as quickly as possible once it has been perpetrated* [5]. When warranty claim is made, it is difficult to prevent the fraud from occurring; the fraudster may try to gain advantage with fictional claims. More importantly, fraud is to be detected before it significantly impacts on the manufacturer's earnings. The emphasis in this thesis is on finding a fraud detection tool for repair warranty claims that identifies fraudulent patterns in the long run. Creating an all-inclusive fraud detection tool is technically impossible. Different perspectives to fraudulent actions re-

quire different tools. Thus it is better to create a system to covering each aspect of fraud separately. SAS, one of the leading service management solution providers, recommends that a complete fraud detection system includes four different controlling approaches [18]:

- Business rules - automatic validation
- Anomaly detection - suspicious data instance recognition
- Advanced analytics - suspicious behavior and pattern recognition
- Social network analysis - analysis of connections of a fraudulent party.

Business rules play the role of fraud prevention from the manufacturer's point of view. The claims are validated using automated rules learnt from previous fraud cases. Anomaly detection tools are then provided to identify anomaly claims. Advanced analytics locates anomalies from a more general perspective, trying to identify the repair partners behaving suspiciously. Finally social network analysis identifies the possibilities of systematic partner fraud. When the results from these four parts are combined, the organization is better equipped to face the threats. Figure 1.1 gives an overview on the warranty claiming process and the relation to fraud control system in the repair claim case.

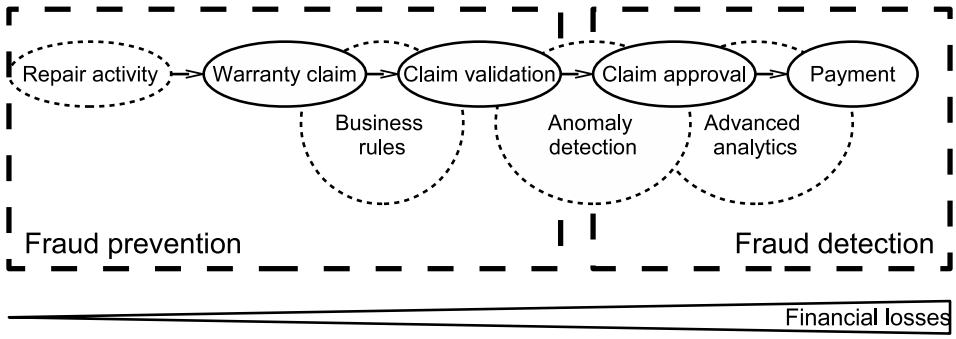

Figure 1.1: Warranty claim process and fraud control. Social network analysis is valid in each point of the figure.

As can be seen from figure 1.1, the sooner fraud is identified, the smaller the manufacturers losses are. Thus business rules are the most cost efficient tool for fraud detection. The identification of new business rules takes place in the advanced analytics section. New fraud patterns and behaviors indicate new rules that need to be implemented. Advanced analytics tools are not the most efficient financially in the present, but important for the future fraud prevention. This thesis studies the possibilities of advanced analytics in warranty

claim area.

Advanced analysis in practice is the use of mathematical tools and techniques. Some of these techniques, clustering for example, can also be implemented in anomaly detection. The difference between these two fraud detection steps is small. Only the objectives differ; anomaly detection seeks single repair anomalies while advanced analytics focuses on suspicious repair partner identification. The tools for the latter case are various, but the basic idea is always similar; compare the recent activity with the historical activity. The comparison is made possible with *profiles*, which represent the historical activities. The methods differ in how the data is presented with the profiles. One possibility for a profile, presented more thoroughly later in this thesis, is to present the data with *signatures* [12]. By definition, signatures are joint probabilities that are built from conditional probability distributions. They present the probabilistic behavior of a repair partner. This profiling method has been used in the literature in anomaly detection. For instance, each new repair activity can be given a fraud score, based on how likely or unlikely the activity is compared to the partners signature [8]. Signatures and fraud detection will be discussed more in section 2.3.1.

1.2 Objectives

This thesis examines the use of signatures in fraud detection. The emphasis will be on identifying behavioral anomalies among the repair partners. In the following text, repair partners will be referred to as *accounts*.

Majority of fraud detection literature concentrates on labeling accounts or data objects as fraudulent or normal. However, the term *fraudulent account* is a strong expression and labeling a repair partner as fraudulent is potentially offensive. A better expression in this context is *suspicious account*, that states that the account is behaving differently than others and should be taken into further inspection. Furthermore, because the data used in this thesis comes from an international consumer electronics company, already the local regulations create anomaly behaviors that the data analyst may not be aware of. These behaviors also need to be pointed out. The purpose of the tool discussed in this thesis is to give the auditors a primary set of accounts where they should start looking for fraud. For advanced analytics purposes, two

different anomaly detection approaches are applied. First, the initialization dataset is analyzed to find accounts already differing from others. Second, a testing dataset is analyzed for behavioral changes.

This thesis answers to the following research questions.

- What are histogram signatures and how do they present the data?
- How can histogram signatures be used in detecting warranty fraud?

1.3 Structure

The process of anomalous behavior detection has multiple steps. The discussion in this thesis, respectively, has been divided into three main topics

1. Profiling - creating histogram signatures
2. Clustering - recognition of anomalies
3. Behavioral outlier detection - peer group analysis

The next chapter presents the literature in the area of warranty claim fraud and how signature based fraud detection can be applied to real data from the literature perspective. Two separate methods are presented; one for the initial outlier account detection and one for the behavioral change detection. The chapter also introduces different measures for signature distance. The third chapter describes the data structure and fraudulent accounts used in the analysis are described. This chapter also presents the signature creation process and discusses the special characteristics of fraud detection methods with signatures. The fourth chapter presents the results of the signature method. Finally, the results of the thesis are summarized and the future areas of application of the proposed method are discussed.

Chapter 2

Anomaly detection methods in fraud applications

Fraud detection is one topic in anomaly detection. Other anomaly detection applications include intrusion, medical information, damage and textual anomaly detection and image processing. Depending on the anomaly detection application, some techniques are more efficient than others. Different anomaly detection techniques include methods such as classification, clustering, nearest neighbor search and statistical analysis [10]. This chapter discusses first the qualities of warranty fraud and then the methods that seem best suited for histogram signatures.

2.1 Warranty fraud

The principle of warranty claim process with repair vendors was discussed in section 1.1. Despite of a good validation system, fraudulent accounts are often able to identify the loopholes in the system and circumvent these automated, often rule-based validators. Estimates of the percentage of fraud costs out of the total warranty costs differ between 10 and 15 percentages [18]. To strengthen the warranty fraud detection system, companies perform both audits and mathematical analysis with the claim data. The accounts may get single fraudulent claims through the system, but what is especially costly for the company is the continuous exploitation of the system. This usually means that the behavior of the account changes after the loophole in the fraud detection system has been found. The fraud detection method should find this

change.

One of the characteristics of warranty claim fraud is that the definition of fraud is quite ambiguous. Most unseen claims are not necessarily fraud, but should raise suspicion. For example, if the repair partner claims a repair done for a device that they have never repaired before, the reason behind this can be that the device has only recently come to the markets in which the repair partner is working. However, it can also be that the device is not in the surrounding market but the repair partner has received information on a device for which they could falsely create claims. These types of situations are rare, but possible. Even if the definition of fraud is fuzzy, there are direct rules that can be created for validation purposes. For instance, there are sets of attributes for each claim that do not allow other attributes. The validation rules such as matching attribute definitions can be based on the device details, for example some cars require a specific engine, and if a repair partner would try to replace the engine with a non-matching engine, the claim would not be validated.

Another characteristic of warranty fraud is that each account can have either very similar behavior as other accounts and their behavior can be completely unique. The difficulty is that both cases may or may not include fraud. For instance, there are networks of repair partners, and if these partners were to perform fraud, their profiles would be very similar, because they would have the same ways of working. At the same time the partners could belong to different markets which would suggest that the repair portfolio should be different. On the other hand, the profiles would be similar also if the environment in which repairs are generated is similar. Furthermore, if the repair partner differs greatly from the others, it may be due to contractual agreements between the manufacturing company and the repair partner and thus would not be an indicator of fraud. There are several cases where differences may arise and it is difficult to say what the reason for the difference is. Despite of all the exceptions, such as those described above, the behavioral difference compared to others is a good starting point for fraud investigation.

The biggest fraud indicator with warranty claims is behavioral change. The identification of this change has been studied in literature, but it has concentrated merely on the behavior change within one account [14, 15]. However, what counts as a behavioral change in warranty fraud field, must be compared

against similarly behaving accounts. New methods have been developed to detect the behavioral change relative to the behavior of an account's peers [5]. One such method, called *peer group analysis*, is discussed in section 2.4.4.

2.2 Fraud detection

Fraud detection techniques can be categorized in several ways. For instance, categorization can be based on input data type or on outlier definition. When the categorization is based on data type, there are three categories; unsupervised, supervised and semi-supervised methods [10].

- *Unsupervised* anomaly detection does not make any assumptions about the data before analysis. It is not known whether the analyzed data instances are already fraudulent or normal. The fraudulent behavior is neither known. The purpose of this data mining is to discover the anomaly patterns, instances or behaviors. Because fraudulent behavior in the warranty claiming process is mostly previously unseen patterns, unsupervised methods are the most useful methods in warranty fraud detection.
- *Supervised* anomaly detection learns from previous fraud. Once they find similar behavior than what is earlier defined as fraudulent, or not normal, an alarm is issued. Most rule-based anomaly detection methods are supervised methods.
- *Semi-supervised* anomaly detection combines the two approaches above. Unsupervised methods are first applied to the data to find the anomalous data instances and supervised learning methods are then taught according to the findings. The whole fraud detection system in warranty claiming works in a semi-supervised manner. The business rules are created through the findings of advances analytics.

Another example of the ways how anomaly detection methods can be categorized is based on the outlier type. An outlier has been defined as

> *an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism [23].*

This definition does not assume anything about the measurement of deviations. Authors have drawn their own conclusions about how deviation is defined and this has created another categorization for outlier detection methods. Devi-

ation has been included in the definition of different outlier types that the anomaly detection methods are able to identify. The most common outlier types include distribution, distance, deviation or density based outliers. Other outlier types include depth, clustering, sub space, support vector and neural network based outliers [26]. The definitions of these outlier types can be found in Appendix A.1. When the different outlier types are put into warranty claim fraud context, some are more suitable than others. The most suitable types are distance, deviation and density based outliers. Distance based outliers are by definition distant from others [27]. Distance calculation method varies, but the outlier scoring is based on the direct distance. For example, if the claiming volume were to be used as the distance measure and if the normal claim volume was low, the account claiming the most would appear as the biggest outlier. Deviation based outliers have characteristics that other accounts do not [2]. If, for instance, an account was reporting all claims in the middle of the night, and the claiming hour was included in the analyzed characteristics, the account would appear abnormal. Finally, density based outliers tend to combine these approaches by considering also the similarities of other accounts [7]. If ten accounts were claiming five to ten claims daily, and other ten accounts reporting 150 to 200 claims with equal variance, these would not appear as outliers, because they are in their own neighborhoods. However, if one account was claiming, say, 50 claims a day, that would appear as an outlier because it would lie far from all the other accounts' groups.

Data and outlier types are characteristics of the outlier detection methods. Different methods can further be divided based on their performance idea. Table 2.1 summarizes the most common anomaly detection methods.

In fraud detection, classification techniques have been highly popular because of the supervised nature. Data objects can be labeled as fraudulent or normal based on previous knowledge of fraud. However, in the context of this thesis, there is no direct knowledge of fraud available. As discussed in the previous section, behavioral change is the best indicator of fraud. Detecting behavioral changes requires the similarity between accounts to be defined and similarity in terms of warranty fraud is complex. Business environment, time of the year or the account contract can determine the behavior of an account and thus creates similarity between certain accounts. This similarity can be measured for instance with claim volumes or claim characteristic distributions.

Table 2.1: Statistical methods in fraud detection [10]

Method	**Idea**
Classification based	Classification of data instances into fraudulent or normal instances
Clustering based	Data objects are clustered and the objects furthest from other objects in clusters are identified as outliers. Also small clusters can be considered as outlying clusters.
Nearest neighbor based	The objects that are furthest from their nearest neighbors are outliers.
Statistical	Outliers are those objects that do not follow a probabilistic model.
Information theoretic	Outliers induce irregular information in the data.
Spectral	Outliers can be seen after projection of data into lower dimensional space.

Similarity measures, fraud complexity, data availability and other characteristics limit the usability of certain anomaly detection methods. Taking a different point of view, the following requirements for warranty fraud detection system can be listed. The method must have

- *Memory.* The history of an account defines the behavior [8].
- *Ability to adapt.* It must detect various anomalies that have not been seen before [14].
- *Ability to learn from others.* Even if the behavior is new for the account itself, it may be normal according to others.

The next sections introduce different unsupervised outlier detection methods that compare the similarities of accounts and try to fulfill the requirements of efficient warranty fraud detection system.

2.3 Profiling

Data processing where the structure of the data is summarized into components with some set of attributes is called profiling [10]. It is important to recognize the correct approach to summarize data. For example, if labeled

data is available and the ways to perform fraud are stable, it is logical to build one common fraud profile and compare the data objects against this. However, when the data is not labeled, one may create a profile for each performer and try to identify possibly fraudulent changes in the profile. Some methods detect several different variables for a profiles behavior, create rules for normal behavior and trace the differences may present fraud. Such a method has been applied for example to track fraud in mobile phone usage [14, 15].

It can be noted that there are various ways to build profiles. In addition to detection of direct variable values, one can use joint probability distributions as account signatures [12]. These signatures can then be applied to fraud detection [8]. This thesis follows these joint probability distribution signatures as profiles.

2.3.1 Signatures

When the values of a variable are discrete, the probability distribution of this variable is often presented using histograms. The probability slots of a histogram are called bins. The benefit of histograms over other tools is the ease of understanding and computation. When timing information is also available, histograms do not prioritize recent transactions over older ones but put the same emphasis on each transaction. Sometimes the loss of timing is seen as a problem. For example, the repair activities in a repair center are likely to change as the consumer needs change. These types of changes are important to notice. This problem can be avoided with a periodic recomputing of the histograms. Another difficulty with histograms occurs if the data is multidimensional and the histograms also need to be multidimensional.

To overcome the multidimensionality problem, joint probability distributions are presented with histograms [12]. This joint probability distribution is called an account's *signature*. Signatures create an individual profile for each account. The ability to present the behavior of an account with probability distributions creates the core of signature methods' efficiency. Furthermore, this ability facilitates the use of signatures in fraud detection. Each new transaction, for which the signatures are built on, can be compared against the historical signature and scored for suspiciousness [8]. The simplicity of histogram signatures is one of their biggest advantages. Before histogram signatures were applied to fraud

detection, the profile for an account was created by determining the normal characteristics of an account [15]. This view is more complex computationally.

Mathematical formulation of histogram signatures is as follows. Let $\mathbf{X}_n = (X_{n,1}, X_{n,2}, ..., X_{n,M})$ be a reported transaction for an account with M different variables at time n. The joint probability distribution for a transaction is defined as

$$P_n(\mathbf{X}_n) = P_n(X_{n,1})P_n(X_{n,2}|X_{n,1})...P_n(X_{n,M}|X_{n,1}, ..., X_{n,M-1}), \qquad (2.3.1)$$

where the conditional probability distributions $P(X|Y)$ are *signature components*. Histograms are a good choice to represent distributions. They lower the amount of information that needs to be stored, because only the sizes of the bins have to be remembered. A benefit of histograms is that they can be used for both categorical and continuous variables. When used with continuous variables, data needs proper discretization.

When histogram signatures are designed, the choice of variables is important, because the histograms are built for each combination of variables. Computational efficiency decreases, if there are too many histograms. The statistical significance of conditioning variables for an account can be tested with a χ^2 test [12]. The test is similar to Pearson's chi-squared test, trying to identify whether the conditioning changes the distribution of the signature variable too much, compared to the distribution without conditioning. The idea is to calculate an expected distribution for each conditioning category and compare the true distribution against it. Only those categories where the count is greater than a pre-defined threshold are included in the distribution comparisons. This ensures the statistical reliability of the test. The χ^2 test statistic is calculated for each account separately. The purpose of the test is to find those variables that are significant for most of the accounts and highly significant for at least some of the accounts. The variables that fill these conditions are then used as the signature components in (2.3.1).

One difficulty with histogram signatures is how to define the set of variables properly and what will be the optimal level for the number of histogram bins K. One approach is to find the number of bins K that maximizes the information entropy (see section 2.3.3) of the histogram. Information entropy measures the level of information within a distribution [40]. If target behavior is available, authors recommend to maximize average weighted Kullback-Leibler measure

(see section 2.3.3) instead of information entropy [12, 8]. Using the average weighted Kullback-Leibler measure the target behavior can be emphasized as much as is required by the analyst. Kullback-Leibler calculates the level of difference in information entropy.

One of the benefits of the histogram method is that we do not have to store all data, but just the histogram probabilities of each account. Adding new data to the histogram probabilities can be done easily by

$$A_n = (1 - w)A_{n-1} + wX_n, \tag{2.3.2}$$

where A_{n-1} is the bin probability after $n-1$ observations and X_n is a vector of ones where the n-th observation belongs to and zero elsewhere. The weighting parameter $0 < w < 1$ can be linked to the number of observations $w_n = \frac{1}{n-1}$, when the signature presents the whole history of transactions. The downside of this is that the behavior of an account may have changed, which is not properly taken into account. Naturally, the behavior change may be an indicator of fraud but can also indicate a change in the claimed product maturity.

To better suite this problem of preferred information, there are other ways to update the histogram. One possibility is to use *exponentially weighted moving average* and fix the weight $w_n = w_{n-1}$ [12, 30, 13]. The closer w is to zero, the less it takes the most recent observations into account and the better the recent history is remembered. If, for example, $w = 0.05$, the observation that occurred ten updating rounds ago, has now only 60% of its original value. To clarify, say some observation has been the only observation in its category, the probability value of this observation is 100%. After ten different observations, however, the probability of this original observation is only 60%. If the number of observations was known, and the new data was not updated, the probability would equal to the relative frequency of the observation. When weighting is used, the number of observations before the last ten observations does not matter.

In addition to the updating process, time information can be seen in signatures as a variable. For instance, if the reporting frequency was to be detected, it is presented as a variable conditioned on other variables. The reporting frequency can also be a conditioning variable with a proper discretization.

2.3.2 Initialization of a signature

When a new account appears, a new signature needs to be initialized. The idea is to use an indexing variable Z that groups the conditioning variables. For instance, if daily warranty claim volumes are grouped with the conditional variable such that $[0 \leq x_1 < 10, 10 \leq x_2 < 20...]$, the indexing variable may have a grouping of $[0 \leq x_1 < 30, 30 \leq x_2 < 60, ...]$. After the first transactions of an account have taken place, the signature histogram can be estimated using the average distribution over all other accounts that are indexed with Z [12].

The importance of indexing by Z is determined with coverage of Z, which is the fraction of customers for whom the indexing is useful, and average value of Z, which is the average distance of indexed and non-indexed signature components. If there are several indexing variables available, the most effective value is chosen as the indexing.

2.3.3 Distances between histograms

Having an account signature raises the question how to measure the difference, or similarity of signatures? Because the behavior of a claiming account depends on several environmental and background factors, the similarity of two accounts must be addressable.

Let P and Q be the signatures of two accounts and $p_k \epsilon P$ and $q_k \epsilon Q$ partition probabilities of these signatures with $k = 1...K$ partitions. A common distance measure used in the literature [5, 44, 10] is the Euclidean distance

$$D_E(P, Q) = \sqrt{\sum_i (p_i - q_i)}. \tag{2.3.3}$$

The Euclidean distance is best suited for continuous variables that follow Gaussian distribution. When calculating the distance over the whole histogram probability distribution, with no assumptions about the shape of the distribution, Euclidean distance is not the best measure.

If the histogram variables are not derived from the Gaussian distribution, the association between these can be calculated with Spearman correlation [33]

$$\rho = \frac{\sum_i (x_i - x)(y_i - y)}{\sqrt{\sum_i (x_i - x)^2 \sum_i (y_i - y)^2}}, \tag{2.3.4}$$

where x_i and y_i are the rankings of i^{th} observation and x and y are the average of these ranks. Because the correlation is calculated for each signature histogram separately, sum of these correlations can be taken as the similarity measure. The problem with Spearman correlation as a distance, however, is that it neglects the actual probability values of the observations.

A popular method for distribution similarity calculations is to use entropy as a baseline for calculations. Several methods are based on *Shannon's entropy* [19, 40, 41]

$$H(P) = -\sum_{k=1}^{K_P} p_k \log p_k, \tag{2.3.5}$$

where p_k is the probability of k^{th} partition and K_P the number of partitions [34]. In multiple dimensions entropy can be summed over the attribute entropies [26, 25].

The actual entropy based distance measure has several modifications. One modification is a mutual information measure between P and Q

$$D_{MI}(P;Q) = -\sum_{i=1}^{K_P} p_i \sum_{j=1}^{K_Q} p(q_j|p_i) \log \frac{p(q_j|p_i)}{q_j}, \tag{2.3.6}$$

where $p(q_j|p_i)$ is conditional probability of Q given P and q_j is the probability of j^{th} partition in Q [1].

As discussed in section 2.3.1, average weighted Kullback-Leibler (AWKL) distance has been popular in histogram signature comparison.

$$D_{AWKL}(P,Q) = -\frac{1}{N}\sum_{i=1}^{N}\left(w\sum_{k=1}^{K} p_{i,k}\log\frac{p_{i,k}}{q_{i,k}} + (1-w)\sum_{k=1}^{K} q_{i,k}\log\frac{q_{i,k}}{p_{i,k}}\right), \tag{2.3.7}$$

where the weighting parameter $0 < w < 1$ can be modified for proper comparison of a distribution against some target distribution. If Q represented a target signature, the selection of w would distinguish between the ability to avoid incorrect grouping ($w = 1$) and the ability to track the individual behavior ($w = 0$) [8].

Taneja [40] discussed optional distance metrics, for example Jensen difference divergence measure also known as Jensen-Shannon divergence or an informa-

tion radius

$$D_{JENSEN}(P,Q) = \sum_{i=1}^{n} \left(\frac{p_i \log p_i + q_i \log q_i}{2} - (\frac{p_i + q_i}{2}) \log (\frac{p_i + q_i}{2}) \right). \quad (2.3.8)$$

That is, Jensen difference divergence presents the *average relative entropy of the source distributions to the average distribution* $\frac{P+Q}{2}$ [35].

Which metric to use, depends on the purpose. AWKL works well with a continuous data that covers all the histogram bins. When the other compared histogram includes empty bins, and the other histogram is complete, Jensen difference divergence gives the best results. It works the best because zero probability does not cause the probability in the other histogram to be forgotten. With other distance measures zero probability makes the total distance zero.

One must note that these entropy based measures are not metrics, because they do not fill the triangle inequality condition $D(P,R) \leq D(P,Q)+D(Q,R)$. However, the square root of Jensen difference divergence measure (2.3.8) fills this condition and is a metric [39]. Thus Jensen difference divergence measure can be used to define the similarity distances between two discrete distributions and use it for example in clustering and nearest neighbor searches [31].

The measures described above calculate the differences between two probability distributions. The difference in (2.3.8) can be generalized to a population of N distributions by

$$D_{JENSEN}(P_1, ...P_N) = \sum_{j=1}^{n} \left(\sum_{i=1}^{N} w_i p_{i,j} \log p_{i,j} - \left(\sum_{i=1}^{N} w_i p_{i,j} \right) \log \sum_{i=1}^{N} w_i p_{i,j} \right), \quad (2.3.9)$$

where $\sum w_i = 1$ [35].

The selection of distance measure depends on the used data. A set of studies have been conducted on the optimum distance measure with varying set of measures and different datasets. The conclusions of these studies have been similar; the selection of the distance measure needs to be adjusted according to the studied data, because different measures are able to identify unlike differences [11, 6].

2.4 Unsupervised fraud detection

In the field of warranty claim fraud detection, labeled data is rare. Unsupervised outlier detection methods are thus a logical choice. Many outlier detection methods can be used in an unsupervised mode. The most common methods are methods that seek to identify data objects that locate in each other's proximity and score the furthest objects as outliers. Clustering and nearest neighbor based methods represent this method category.

Clustering and nearest neighbor search are similar methods. Both depend on the performance of the distance measure, implying that the choice of data preprocessing technique is important. The importance is only increased with categorical data [3]. The difference of these two method types is that clustering techniques calculate the distance of the data instance with the whole cluster while nearest neighbor methods measure the distance with the local neighborhood of the data instance [10]. It is difficult to state which of the approaches is more correct and both of the methods should be examined with the data available.

Both techniques have problems. One problem with clustering as an outlier detection method is that they were originally developed to identify clusters, not outliers. The recent shift in focus towards more data driven outlier detection methods has also driven the development of cluster based outlier detection methods. The problem with nearest neighbors is that if the data points have differing number of neighbors, the method fails to identify the outliers or label normal data instances incorrectly as outliers.

2.4.1 Clustering

There are three main clustering methods that several articles have concentrated on. These are k-means, k-medoids, and hierarchical clustering techniques. The basic idea in k-means and k-medoids clustering is to guess an initial set of k cluster centers and apply a minimization algorithm to the data until a global minimum of total distances from the cluster centers is minimized. The difference between k-means and k-medoids is that k-means uses Euclidean distance as object distance while with k-medoids the object distances can be determined by any dissimilarity measure. Hierarchical clustering differs from these methods in the initial approach. The clusters are created step by step,

first combining the most similar data objects. Clustering ends when either the wanted number of clusters is found or the smallest distance left exceed a determined threshold [22].

The optimal clustering to k clusters is difficult. Optimality of clustering can be evaluated for instance by partitioning the continuous variables to create discrete distributions and then use entropy to compare these partitions. Optimal clustering based on this partitioning is able to detect different types and sizes of clusters [17]. However, this does not consider the optimal number of clusters. Literature includes several measures to compare clusterings with different numbers of clusters, such as Bayesian information criteria, F-statistics, coefficient of separation and partition coefficient. Each measure has its own application areas. Many measures are applied to probabilistic clustering, where the number of parameters is an important issue. One measure that does not focus on the number of parameters and is thus well applicable to this thesis is **silhouette coefficient**. It fits well to distance based clusters. It compares the average distance of a point x to other members in its cluster C, $a(x) = \frac{1}{|C|-1} \sum_{y \epsilon C} d(x, y)$ with the average distance to members in second best fitting cluster G, $b = min_{G \neq C}(\frac{1}{|G|} \sum_{y \epsilon G} d(x, y))$. The average of silhouette coefficient $s(x) = \frac{b(x)-a(x)}{max\{a(x),b(x)\}}$ indicates how accurate the clustering is.

Clustering as an outlier detection method has three basic approaches how to distinguish between normal data and outliers. First, the normal data instances are expected to belong to a cluster. Data objects that do not belong to any clusters are outliers. Second, the normal data instances are expected to lie close to the cluster centroid and outliers locate on the edges of the clusters. Third, normal data builds large and dense clusters while outliers have only few partners in their cluster [10]. Each of these approaches may raise suspicion on different accounts. It needs to be examined which approach finds the best outliers in warranty claim data.

2.4.2 Nearest neighbors

The basic idea of nearest neighbor analysis as an outlier detection method is to calculate the distance of data objects from their closest neighbors and point out the objects that lay far from their neighbors. The efficiency of such methods depends on the applied similarity measure.

In addition to the choice of similarity measure, the performance of nearest neighbor methods depends on the selection criteria of neighbors. The neighbors can be selected based on two separate parameters, the number of neighbors or the distance from the data object. Methods based on the number of neighbors calculate the outlier score according to the average distance. Methods based on the distance criteria count the neighbors locating closer than the criteria and scores the data objects as outliers if the neighborhood is scarce. These method criteria often fail to identify outliers if the data density varies. To overcome this problem, some methods have been developed to consider also the density of the neighborhood. One density based method is introduced in the next section.

2.4.3 Local outliers

Local outliers locate further from their closest neighbors than an average neighbor of those neighbors. Each data object is given an outlier score called local outlier factor (LOF) [7]. To introduce this measure, a few helping calculations need to be introduced. $K\text{-}distance(p)$ calculates the distance to the k-th nearest observation, where k is determined by the analyzer.

$$k\text{-}distance(p) = d(p,o)\ s.t. \begin{cases} \text{for at most } k\ o'\epsilon D\backslash\{p\}\ d(p,o') \leq d(p,o) \\ \text{for at most } k\text{-}1\ o'\epsilon D\backslash\{p\}\ d(p,o') < d(p,o), \end{cases} \tag{2.4.1}$$

where D is a set of data objects. *K-distance neighborhood* includes the observations that are among the k nearest neighbors.

$$N_{k\text{-}distance(p)}(p) = \{q\ \epsilon\ D|\ d(p,q) \leq k\text{-}distance(p)\} \tag{2.4.2}$$

Reachability distance identifies whether an observation, for which the outlier factor is calculated, also belongs to the K-distance neighborhood of its nearest neighbors. Furthermore, *local reachability density* defines the density of observations for which the previous statement is true.

$$reach\text{-}dist_k(p,o) = max\{k\text{-}distance(o), d(p,o)\} \tag{2.4.3}$$

$$lrd_k(p) = \left(\frac{\sum_{o\epsilon N_k(p)} reach\text{-}dist_k(p,i)}{|N_k(p)|} \right)^{-1}, \tag{2.4.4}$$

where $|N_k(p)|$ is the number of observations in the neighborhood, i.e. k.

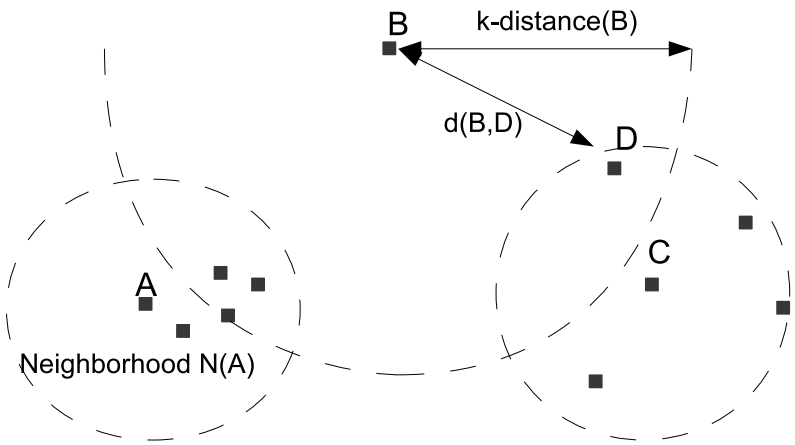

Figure 2.1: Illustration of local outlier factor calculation.

Finally, local outlier factor can be calculated using these results.

$$LOF_k(p) = \frac{\sum_{o \epsilon N_k(p)} \frac{lrd_k(o)}{lrd_k(p)}}{|N_k(p)|} \tag{2.4.5}$$

Reach-dist$_k(p, o)$ decreases the statistical fluctuations of local outlier factor when $n > 2$. As can be seen from the definition of LOF, the performance of LOF relates closely to the number of neighbors taken into account. LOF aims at identifying density based outliers efficiently.

Looking at figure 2.1, if the local outlier factor is calculated considering four closest peers, data objects A and C both have their own clear neighborhoods. Those neighborhoods do not have the same density, but the all the members are each other's' closest neighbors, and will get LOF value 1. The neighborhood for object B has greater distances from the object B. When the local outlier factor for object B is calculated, *local reachability density* finds out the neighborhood densities of the closest neighbors, including object D and the objects in the neighborhood of object A. Because these objects in the neighborhood of B do not have object B in their neighborhood, object B will get high local outlier factor value.

The advantage of local outlier factor is that it considers only the local differences of data instances. However, this can be a disadvantage as well. For example, if the closest data instances of an outlier are also outliers and locat-

ing far from all the others, LOF-score would remain low. Thus the choice of number of neighbors included in the LOF calculation is important.

2.4.4 Behavioral change analysis

Clustering and nearest neighbor analysis methods are the best in data mining tasks where the data set does not evolve over time. However, in many applications, especially in fraud detection, the development over time is an important influencer. Similar to nearest neighbor techniques, peer group analysis compare an account against its nearest neighbors and further track the development of their accounts and score the accounts according to their changes [4]. This method is unsupervised outlier detection method because it does not require labels in the data but focuses on comparing the changes within an account. Break point analysis, on the other hand, detects the within account changes and scores the accounts based on the historical changes within the account [5]. Activity monitoring is similar to break point analysis, giving anomaly scores for each data instance based on its suspicious attributes [15]. The suspiciousness of an attribute needs to be learnt from historical data. Although proven very useful in credit card fraud detection, activity monitoring is not applicable in this system, because of the lack of knowledge in suspicious behavior. Peer group analysis on the other hand considers the relation to neighbor accounts and is valuable in the context of this thesis.

Peer group analysis is based on signatures. The type of the signature is not important for the performance of the method. The original article studied signatures based on single values of different variables, but there is no reason why the signatures introduced in section 2.3.1 could not be used. The idea to detect behavioral changes in accounts differs from clustering methods such that they do not try to group the whole data to clusters but compare each account individually against the closest similar accounts. Thus the initial set up for the outlier detection requires determining of the nearest neighbors. Because the variance within the neighborhood is relevant, the original authors of peer group analysis used direct distance to determine a first set of nearest neighbors and then further selected those nearest neighbors j that had the least different variance ($VarDiff$) from the target account A_i

$$VarDiff(A_i, A_j) = (var(A_i) - var(A_j))(var(A_i) - var(A_j))' \qquad (2.4.6)$$

The accuracy of clustering methods depends often on the number of observations in the cluster. Similarly, peer group analysis performs differently when the number of peers varies. Furthermore, the performance of this method depends on the choice of distance measure. Similarly to clustering methods, with continuous variables Euclidean distance and variance are the first measure choices to find the peers for accounts. Signatures based on histogram probabilities require some other distance measure. The proper distance measure was discussed in section 2.3.3.

The actual outlier detection method in peer group analysis is simple. Peer group defines the target behavior T, and if the behavior $A_{i,j}$ of an account j at time i deviates greatly from this target, it is flagged as an outlier. Three measures can be used as the average performance measure of the group. These measures are centroid, trimmed centroid and medoid. For instance, centroid $T_{i,j}$ is defined

$$T_i = \frac{1}{npeer} \sum_{j=1}^{npeer} A_{i,j}. \tag{2.4.7}$$

Peer group analysis also considers the dispersion of accounts over time

$$V_i = \frac{1}{npeer - 1} \sum_{j=1}^{npeer} (A_{i,j} - T_i)(A_{i,j} - T_i)'. \tag{2.4.8}$$

The flagging score for each account is determined with the variance V_i and group performer T_i

$$C_{i,j} = (\mathbf{A}_{i,j} - \mathbf{T}_i)'(\mathbf{V}_i)^{-1}(\mathbf{A}_{i,j} - \mathbf{T}_i). \tag{2.4.9}$$

Peer group analysis requires quite large dataset to perform efficiently [4]. The size of the dataset may vary quite a lot, depending on the level of detailed information. Occasionally it is more relevant to study a smaller sample of similar accounts to detect the most deviating accounts. Because of this, one peer group size does not engage all the needed details to classify outliers from normal data.

The performance of peer group analysis has been studied in credit card fraud. Results show that peer group analysis is very promising in detecting suspicious accounts, and should be a part of an efficient fraud detecting system. Moreover, results imply that peer groups that are built with three months of credit card

data is good to detect fraud, but a robust peer group building time cannot be estimated [44]. The timeframe used for building the account signatures in warranty fraud context thus needs to be studied carefully.

2.5 Detection method performance metrics

The outcome from a fraud detection method is an outlier score for each data object. Objects with highest score are then defined fraudulent. When true labels can be determined for the data objects, one can calculate the performance of the method with different performance measures. Percentages of correctly identified fraud cases (TP), correctly identified normal cases (TN), falsely identified normal cases (FF) and falsely identified normal cases (FN) are common performance measures [10]. More advanced measures calculate ratios between falsely and correctly labeled data objects. For instance, the receiver operating characteristics curve (ROC-curve) shows the trade-off between TP and FF percentages [9]. However, these metrics are applicable only if fraudulent and normal accounts can be identified. When labels are not available, the performance of the method needs to be clarified from the data by using case by case outlier analysis.

Achieving a correct outlier score for data objects is the goal of the detection methods. When several types of outliers exist, and also several types of outliers appear, the scoring becomes more difficult. Several sources in the literature concentrate on methods that discover only one type of outliers. However, one option for outlier scoring is to use a combinatorial model, where several local outliers are first individually observed and later combined to global outliers [36].

Chapter 3

Preparation of account signatures for fraud detection

3.1 Data structure

The data set used in this thesis contains warranty claim data from a consumer electronics company. It consists of a large number of rows, $N > 2000000$. Each row contains information variables, such as claiming date and account information, and a number of categorical attributes $x_{i,j,k}$. In this thesis, three attributes have been selected randomly to represent the claim details. Each data instance $\mathbf{X}_{i,j}$ for account i at time j is defined by these three categorical variables. Categorical variables have limitations in what combinations they can have. There are 7 different categories for X_1, 20 categories for X_2 and 52 categories for variable X_3. Furthermore, there are only 124 categories when the data is categorized based on variables X_1 and X_2. If the absolute possible values of X_1 and X_2 are multiplied, which equals to 140, it can be seen that there are 16 category combinations that are not valid. Similarly, the combination of variables X_2 and X_3 can have 320 different combinations and finally categorizations over all three variables reaches 1535 combinations of variables. The numbers of possible variable combinations can be seen in table 3.1.

The numbers of combinations are smaller than the product of the number of absolute categories because the repairing process only allows certain repairing features to be combined. Other combinations are automatically declined when the claim is sent for validation. When the correct combination of variables is selected, the data is divided into probability bins described in section 2.3.1.

Table 3.1: Variable category summary. $X_i|X_j$ means that the data is divided into categories by X_i conditioned on X_2.

Variable combination	**Nr of categories**
X_1	7
X_2	20
X_3	52
$X_2\|X_1$	124
$X_3\|X_2$	320
$X_3\|X_2\|X_1$	1365

The following example clarifies the situation.

3.1.1 Data example

Table 3.2: An invented example of the data structure for one account.

Day	**Cost**	**Difficulty**	**Duration**	**Repairs**
1	Cheap	Easy	Week	1
1	Cheap	Easy	3 Hours	3
1	Expensive	Easy	Hour	2
1	Expensive	Difficult	3 Hours	1
2	Cheap	Easy	Month	2
2	Cheap	Easy	Day	1
2	Cheap	Easy	3 Hours	2
2	Expensive	Easy	Hour	1
3	Cheap	Easy	Month	2
3	Cheap	Easy	Day	3
3	Cheap	Easy	Week	1
3	Expensive	Difficult	3 Hours	1
4	Cheap	Easy	Week	1
4	Cheap	Easy	3 Hours	4

Each repair can be assumed to have 3 parameters; cost, difficulty and duration. The example dimensions are purely hypothetical. The letters in brackets below are linked to figure 3.1; they present the value of the category.

- The repair is either cheap (C) or expensive (E)

- There are easy (e) and difficult repairs (d). Only expensive repairs can be difficult.
- The customer can decide when he or she wants the repair to be ready. Cheap repair can last 3 hours ($3h$), a day (da), a week (w) or a month (m) and expensive repair lasts either one hour (h) if it is easy or three hours if it is difficult.

Table 3.2 shows the example data structure. Data covers four days of data and is presented for one account only. The last column shows the summarized repair volume for each combination of categories.

Table 3.3 presents the categorical combinations in this example similarly to table 3.1. These combinations could be available only for one account, while other accounts can do also expensive repairs that last for a week for example. When the signatures are compared against each other, the missing combinations are those that may be contributing to the distance the most. The next sections present how the signatures are created from the dataset.

Table 3.3: Variable category summary for the example.

Variable combination	**Nr of categories**
Cost	2
Difficulty	2
Duration	5
Difficulty | **Cost**	3
Duration | **Difficulty**	6
Duration | **Difficulty** | **Cost**	6

3.2 Signatures

The data set includes three categorical variables and one calculated variable. The signature for account i defined in equation 2.3.1 is stated as

$$P_i(\mathbf{X}_i) = P_i(X_{i,1})P_i(X_{i,2}|X_{i,1})P_i(X_{i,3}|X_{i,1}, X_{i,2})P_i(A_i|X_{i,3}, X_{i,1}, X_{i,2}). \quad (3.2.1)$$

Let $p_{k,l,m}$ be the histogram for k^{th} category of X_1, l^{th} category of X_2 and m^{th} category of X_3. Figure 3.1 shows the idea of a signature for one account. The figure presents the data in the example data above. There are in total 25 repairs. 80% of them are cheap, 20% expensive ($P(X_1)$). This is presented

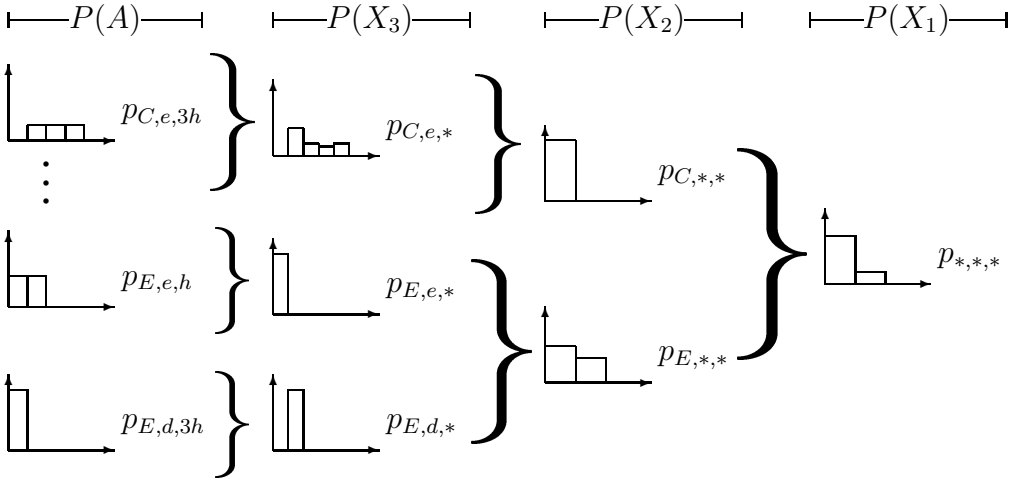

Figure 3.1: Signature. Histogram $p_{*,*,*}$ represents the histogram of variable X_1 (Cost) when no conditionalization has been made. $p_{E,e,h}$ on the other hand presents the conditional histogram of daily repair volume when $X_1 = Expensive$, $X_2 = Easy$ and $X_3 = One\ hour$.

as histogram $p_{*,*,*}$. Furthermore, 100% of the cheap repairs are easy and 45% of those are done within 3 hours, showing that the shortest repair time is the most preferred. Analysis on these repairs shows that either 2, 3 or 4 repairs are usually performed each day.

It is important to understand the structure behind the signature because fraud can be detected within one volume histogram ($P(A)$) or within some categorical histogram, and the changes that reveal the fraud should be monitored. Furthermore, the use of a cumulative variable such as daily claim volume creates an issue in the signature creation process. Conditioning with more variables splits the volumes to smaller counts and this means that the volume histograms need to be set every time a conditional variable is added. Figure 3.2 gives a graphical example of the component $P(A|X_1)$ of an average signature calculated over the whole dataset. The figure only shows the histograms conditioned over only one variable with seven attributes used in the empirical results section.

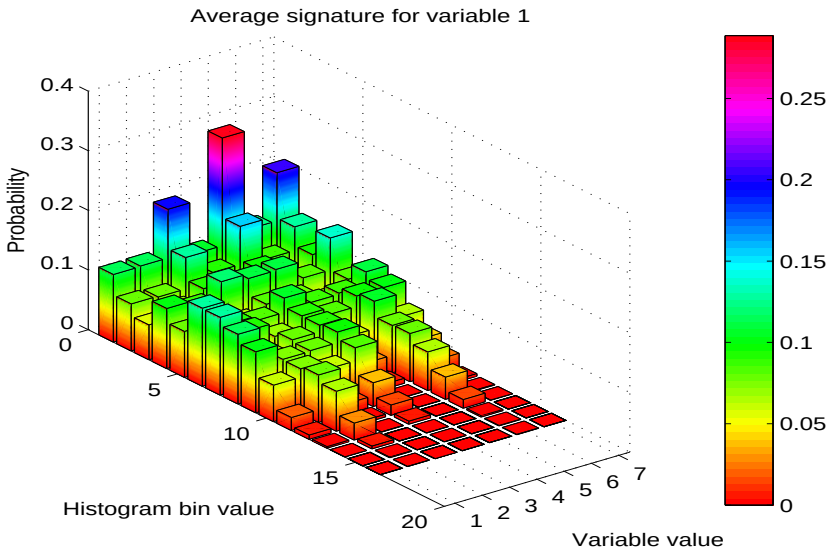

Figure 3.2: Average signature. Histograms present the conditional probabilities of daily volumes, conditioned on one variable $P(A|V_1)$.

3.3 Signature distance

Based on a review of the data, the Jensen difference divergence measure (2.3.8) was chosen as the primary distance measure. Creation of the signatures showed that several signatures in fact do not fill all the signature components. Thus zero probabilities are an important part of the signature and the distance measure must treat zero values correctly. Jensen measure considers the zero values more correctly than other measures.

When the histogram distribution is based on daily claim volumes the fraud analyst may wish to weight the high volumes more than the low volumes. This aims to detect the most expensive changes from the manufacturer's point of view. As a distance measure, one can use the square root of weighted Jensen difference divergence measure

$$D_{WJENSEN}(P,Q) = \sqrt{\sum_{i=1}^{n} \alpha_i \left(\frac{p_i \log p_i + q_i \log q_i}{2} - (\frac{p_i + q_i}{2}) \log (\frac{p_i + q_i}{2}) \right)}, \tag{3.3.1}$$

where $\sum_i \alpha_i = 1$.

The Jensen measure only gives the dissimilarity measure for one histogram.

Because the outlying behavior can be revealed in different signature components, an outlier score should be calculated for each component separately. Using the weighted Jensen difference divergence measure the difference of signatures Y_1 and Y_2 in component A is

$$D_A(Y_1, Y_2) = \sum_{i=1}^{K} D_{WJENSEN}(Y_{1,i}, Y_{2,i}), \tag{3.3.2}$$

where K is the number of histograms in component A. The total distance of signatures D_{tot} is then calculated as

$$D_{tot}(Y_1, Y_2) = \sum_i \beta_i D_i(Y_{1,i}, Y_{2,i}),\ i = \{X_1, X_2, X_3, A\}. \tag{3.3.3}$$

The weighting parameter β is needed because the component distances (3.3.3) are not comparable. The number of bins in histograms varies in different components and thus summing in signature distances in equation (3.3.1) gives different values for different components. Because the nature of fraud is not known, each signature component should be given the same emphasis. The parameter β is estimated using the average of component-wise distances

$$\frac{1}{\beta_A} = \sum_{i=1}^{N}\sum_{j=1}^{N} D_A(Y_i, Y_j)/N^2, \tag{3.3.4}$$

where N is the number of accounts. Another way to calculate the total distance is to normalize the distances.

3.4 Modifications to fraud detection with signatures

As discussed in section 2.4.1, the most common metric in clustering is the Euclidean distance. Because Euclidean distance is not usable with histogram signatures (2.3.3), weighted Jensen difference divergence (3.3.1) measure is used instead of the Euclidean distance.

Weighted Jensen difference divergence measure can replace Euclidean distance also in peer group analysis. In literature, variance and distance of accounts determine the peer groups. Variance and entropy both can be considered as a measure of uncertainty. Similarly, entropy was defined by Shannon as a measure of uncertainty [43].

Peer group statistics (2.4.7) and (2.4.8) in section 2.4.4 define the behavior of a peer group. For signatures, the variable defining the behavior is the same as the signature probability $A_{i,j} = P(A_{i,j}|Y)$, where Y is a set of conditioning variables. Because the preferred difference measure is weighted Jensen difference divergence measure, a better group variance for the group of X_j is

$$V_i = \frac{1}{npeer - 1} \sum_{j=1}^{npeer} D_{WJENSEN}(X_j, T_i), \tag{3.4.1}$$

and the outlier score thus becomes

$$C_i = \mathbf{D}V_i^{-1}\mathbf{D} \tag{3.4.2}$$

Equation 3.4.2 gives the suspicion scores for each updated day. Because this score includes also the distance from peers, this distance easily overrules the changes happening in the group and the member that has the highest distance in the beginning of the peer group analysis remains to have the highest score. Removing the initial distances from the account scores gives a better indicator on the actual changes. Thus the final outlier score is

$$C_i = \mathbf{D}V_i^{-1}\mathbf{D} - D_0 \tag{3.4.3}$$

3.5 Outlier summary and performance measure

Local outlier factor method and peer group as fraud detection methods require a proper guess of the number of peers used. However, when there is no knowledge of how many peers should be compared, clustering seems the most intuitive choice. When clustering is used for outlier detection, a common method is to calculate the score as the distance to the cluster center 2.4.1.

Because of the data structure (section 3.1) and the type of fraud (section 3.6), it is difficult to bring out the most important outliers. Some forms of fraud might be caused by the differences only in the distribution of one variable, while other forms of fraud only appear in the most specific level of categorization. The best way to make sure that all fraud types are detected is to calculate a suspicion score on each variable level.

The outlier identification system requires several parameters that the user has to define. These parameters are based on environmental factors. Table 3.4 summarizes the parameters used in this system.

Table 3.4: System parameters

Parameter type	**Parameter**	**Explanation**
Weight parameters	α	Histogram information weight
	β	Component weight
	w	Group divergence weight
Group parameters	k	Number of nearest neighbors
	N	Number of clusters

Because the purpose of outlier detection is to find the most suspicious accounts the rank $r_{i,n}$ of the created fraudulent account X_i is used as a performance measure. The smaller the rank, the sooner the account will be taken under fraud investigation and the monetary losses that increase over time are minimized.

3.6 Fraud indicators in warranty fraud

The method used in fraud detection needs to track several types of suspicious indicators. The performance is estimated based on the ability to discover all of them. The following fraud indicators need to be covered.

- Abnormal behavior
  - High volume - Account transactions are higher than for peer accounts
  - Varying focus - Account transactions do not distribute similarly than peer accounts
- Behavioral change
  - Volume change - Account transaction volumes increase from normal level
  - Focus change - Account transactions
- Single outliers
  - Sudden high values - Account transactions are expected but surprisingly high in volume

– Unobserved value - Combination of variables is unknown from the past

Term such as "normal level", is not easy to determine. Normal level suggests that the level is calculated from some data where there is no fraudulent activity involved. But because the fraudulence of the data cannot be confirmed, the uncertainty that the fraudulent data points bring needs to be accepted. Furthermore, there are several ways to define the normal level, for instance, one can use the medoid or centroid as a normal reference [4]. In short, the comparison depends greatly on the used definition of similarity.

The tested fraud accounts are created as a combination of real accounts and simulate the outlying behavior. This way we can also estimate the performance of the similarity measurement. The creation of the tested accounts is explained in more detail in section 4.2 and Appendix A.2.

Chapter 4

Performance of methods

This section presents studied methods with real data. First, the signature initialization process is discussed step-by-step. Second, signatures are tested with clustering and local outlier methods. Finally, peer group analysis is performed using signatures. The data collecting and structuring before initialization is handled with Qlikview 9.0 [37]. The main analysis; signature creation, outlier detection and finally peer group analysis are performed with MATLAB [38].

4.1 Signature initialization

4.1.1 Dataset

The pre-modifications to the data were described in chapter 3. The number of conditioning variables was limited to three variables. Furthermore, there was only a limited amount of data available for signature initialization and clustering, and then further to proof the performance of peer group analysis. Table 4.1 shows the dataset information available for this thesis. Learning data refers to the signature initialization and clustering data, and testing data refers to the peer group analysis data.

Table 4.1: Dataset for outlier detection

Variable	**Learning data**	**Testing data**
Number of days	141	45
Number of accounts	73	65
Number of transactions	2 062 475	509 972

As can be seen from table 4.1, some accounts did not contribute to the testing data. These accounts most likely are quite small in volume and can be kept in the learning data without endangering the data quality. In addition, the dataset was modified according to table A.1 to assure that there are outlying behavior in the data.

When analyzing the data, logarithmic values of claim volumes were used. The distribution of logarithmic claim volumes is close to normal distribution, which enables more reliable analysis results than using plain volumes.

4.1.2 Creating the signature

The first step in creating the signature is to find a proper categorization and histogram bin count for this categorization. As discussed in section 2.3.1, the optimal histogram for a dataset maximizes the AWKL-distance (2.3.7) over all accounts and against some target behavior. Because there is no fraudulent target data available, an average of the whole dataset is assumed as the target behavior. The weighting for AWKL-distance should be balanced so that individual behavior is identified without losing the grouping information (section 2.3.3), meaning that the weighting should be close to 0.5. The number of bins in the optimal histogram also limits the signature creation. The higher the bin count, the more computationally demanding the analysis. Another point to be addressed is how and with which variables the data is categorized. If the categorization does not bring any additional information or is inaccurate, categorization is useless.

Let the first categorization be done with variable X_1. Figure 4.1 shows how the bin count of the optimal histogram increases when the weighting in AWKL-distance increases. The possible bin count has been limited to hundred bins. Figure shows that with weight 0.4 the optimal bin count is 16. When $w > 0.5$, the optimal bin count is so high that including this in the analysis would be very inefficient. Based on this figure, weight 0.4 would be chosen as the AWKL weighting. The optimal bin count is calculated over a set of histograms. The left graph of figure 4.2 shows how the histogram with bin count 16 reaches a little higher information entropy value than other histogram bin counts. When calculating the histograms, bin sizes of equal widths has been assumed.

The next step in signature creation is to determine the proper set of condi-

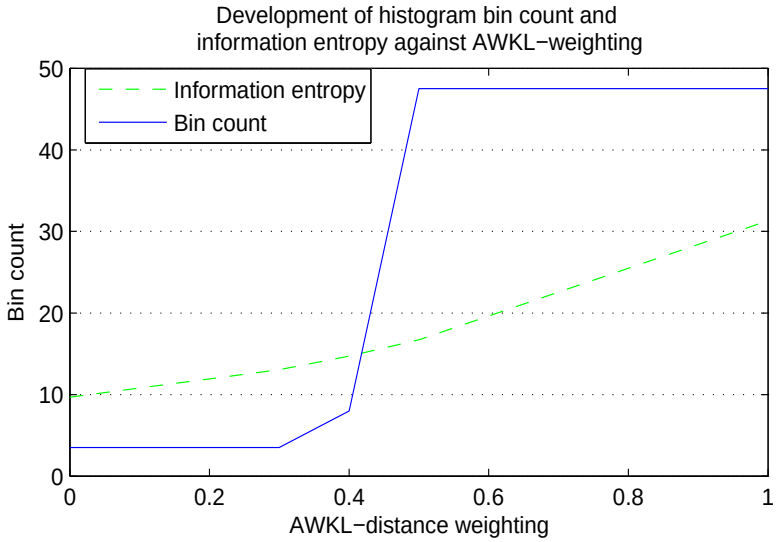

Figure 4.1: Development of the bin count and information entropy against weighting parameter w.

tioning variables for analysis. This process was explained in section 2.3.1. The problem of using volumes when detecting outliers is that the same histograms that were employed when examining the total volumes cannot be used, because the daily volumes are scattered over the conditioning variables. Thus the optimal histogram bin count needs to be determined for each combination of variables separately.

Table 4.2 shows the χ^2 test results for each combination of variables. First, each conditioning variable was tested separately. If each variable was proven significant, the one with the best test results was taken as the first conditioner and another conditioning variable was included in the analysis. If the combinations of variables were proven significant, the data was analyzed with all three conditioning variables. The χ^2 test results show that the best level of information is achieved when the data is conditioned with all the possible variables. The results from the significance test are good. As can be seen also from the results, when the conditioning happens over all the variables, the order of variables is not relevant, because the same histograms are built in every case.

Variable combination that has the best results and still meets the significance requirements is chosen as the final combination. Table 4.2 summarizes the results.

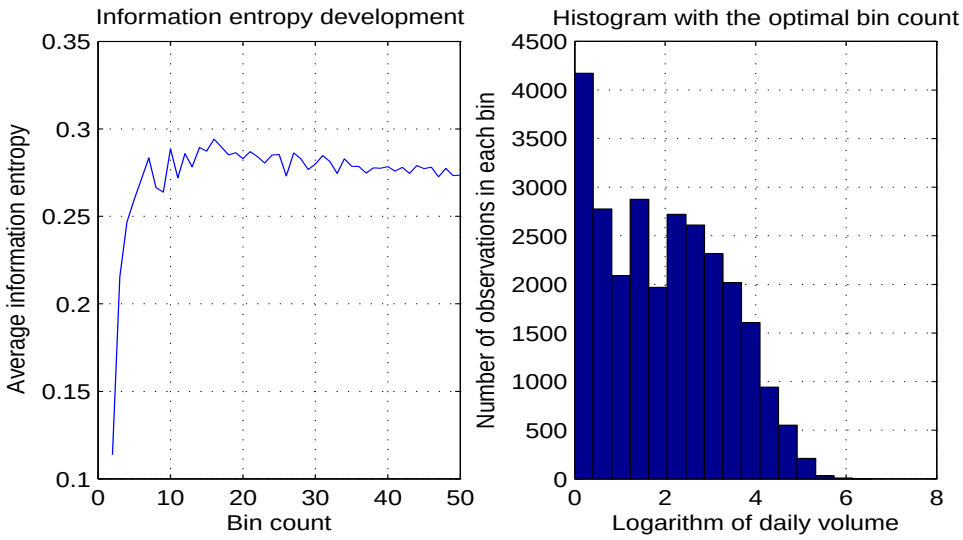

Figure 4.2: Optimized bin count for daily histograms with variable X_1

Figure 4.3 shows the proportion of p-values. It shows the percentage of accounts that pass the test with a specific p-value. X-axis in the figure has a square root -scale. The figure shows that for majority of accounts the conditioning is very significant but the increase in significance level is quite slow after that. Development for the p-value was similar to all the combinations of conditioning variables.

When the conditioning is done with all the variables, the optimal bin count is obtained with AWKL-weight 0.6. The development of the weighting against the maximal information entropy is shown in figure B.1. The optimal bin count is now 11 (left graph in figure 4.4). It should be noted, that the histogram with this set up is highly skewed towards the low values of claiming volumes (right graph in figure 4.4), just like was discussed above.

4.2 Clustering and nearest neighbor methods

Two additional accounts were created to test the performance of different clustering and anomaly detection methods. These anomalous accounts were created to represent two approaches, signatures with suspiciously high volumes and signatures where the distribution of repairs is abnormal i.e. the account has abnormal behavior (section 3.6). More detailed explanations of the creation of these accounts is in appendix A.2.

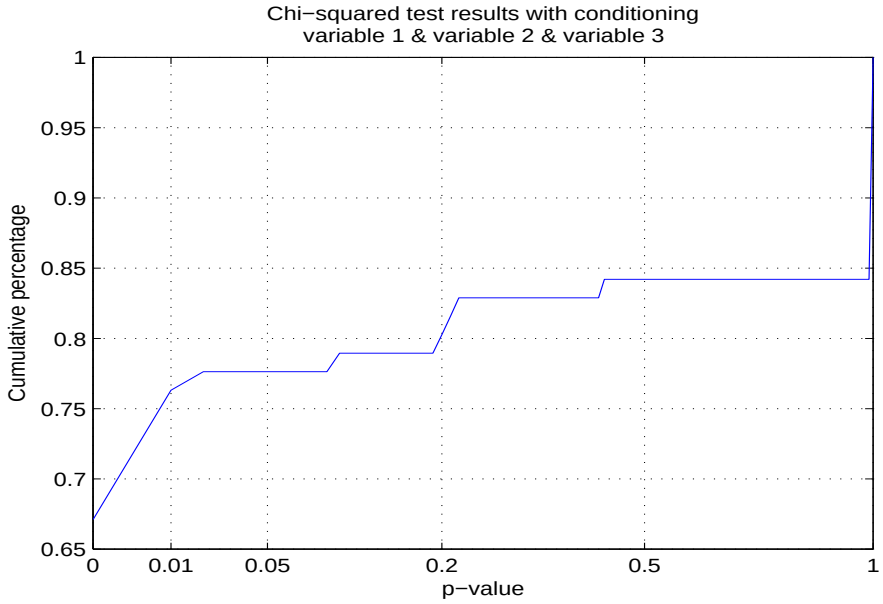

Figure 4.3: p-values for the account set with all three variables. 95% significance level was used for the χ^2 test.

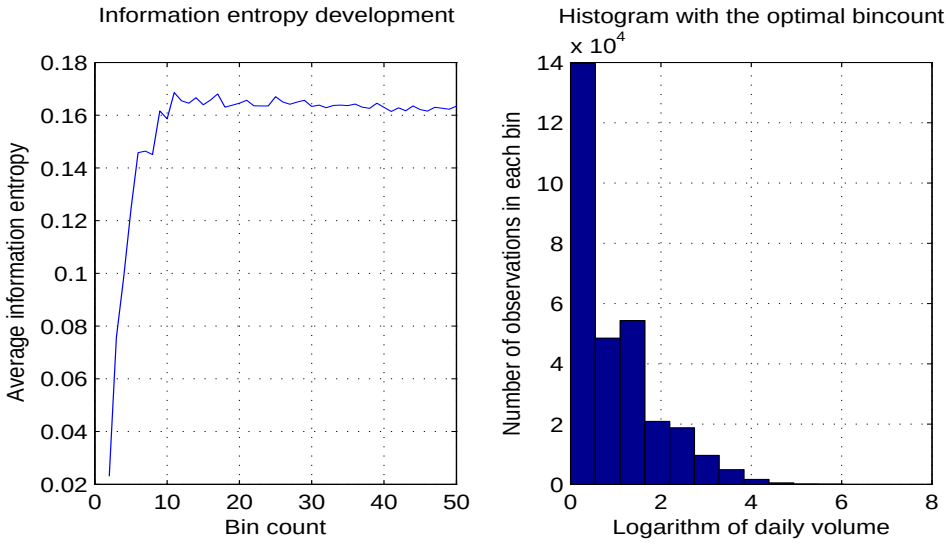

Figure 4.4: Optimized bin count for daily histograms with all variables

Table 4.2: Percentage of accounts for which the χ^2 test results were significant (< 0.05) or highly significant (< 0.01). Test results are calculated for each variable combination separately.

Variable	**<0.01 significance**	**<0.05 significance**
X_1	67.1%	71.2%
X_2	66.7%	68.0%
X_3	76.7%	76.7%
$X_3 \mid X_1$	78.1%	78.1%
$X_3 \mid X_2$	76.7%	76.7%
$X_3 \mid X_1 \mid X_2$	72.6%	76.7%
$X_1 \mid X_2 \mid X_3$	72.6%	76.7%

Before scoring the outliers with clustering method, the clusters need to be created. The number of clusters has a great impact on the efficiency of clustering. The number of clusters should be such that the cluster structure minimizes the lost information. Figure 4.5 shows how the summarized and average distance of data objects develops when the number of clusters increases. Distances of clusters are based on the Jensen difference divergence measure (3.3.3). Silhouette coefficient, which was discussed in section 2.4.1, reaches its maximum already with two cluster. As can be seen from the figure, the decrease of average distance from cluster centers is the fastest when there are only two clusters. The analysis will be based on two clusters.

Figure 4.6 presents the clusters. The account distances are scaled to two dimensions to visualize the distances in a graphical form. The outlying accounts that were created for performance analysis are highlighted in the figure. A ready-made clustering function in MATLAB called *kmeans* (see MATLAB help in [32]) was used to identify the clusters. The figure shows that the clustering method does not find clusters that would be easily discovered. On the contrary, the clusters are mostly selected based on the horizontal position in the figure.

4.2.1 Outlier detection

One way to score the accounts is to order the accounts by the distance from others in the same cluster. Figure 4.7 presents these distances for each account. The greater the distance is from others, the more suspicious is the account. As

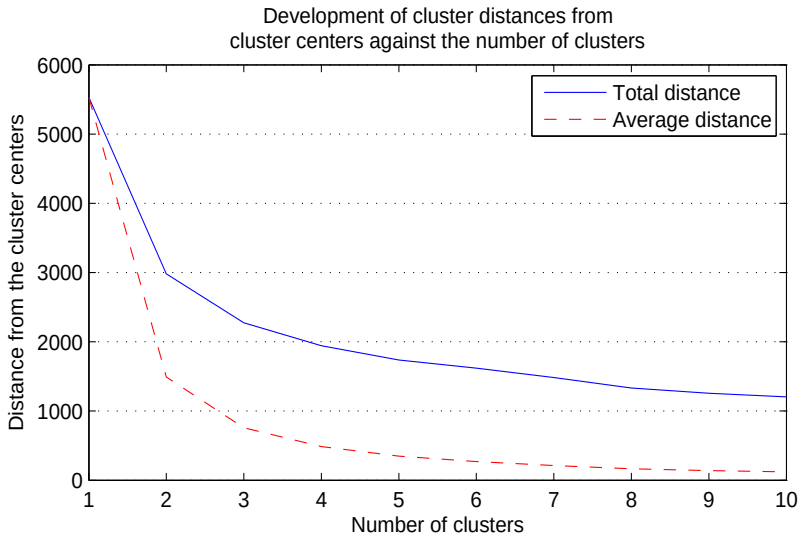

Figure 4.5: Average and sum of distances of cluster members as a function of cluster count.

can be seen from the figure, the created anomalous accounts are observed quite well by the clustering method, i.e. they locate far from their cluster center. The behavioral outlier is detected as the most suspicious case and the account with high repair volume is identified as the 15th most suspicious.

As discussed in section 2.4.3, local outlier factor considers the local differences of accounts better than clustering and should be able to identify the created outliers more efficiently. Figure 4.8 shows the local outlier factors of the accounts. The distances of accounts have been calculated with equation (3.3.3), similarly to the clustering method. According to literature, the closer the score is to one, the more normal the observation. Again, the account created with abnormal repair volume receives a high score. This time it is identified as the ninth biggest outlier. This behavioral outlier is the most suspicious account. The results of local outlier factor analysis can be understood better by looking at the figure 4.6. The account with abnormal behavior locates far from all its close neighbors and thus receives a high score. Account with abnormal volume locates in a neighborhood of two other accounts but the next closest accounts are relatively far. The results were calculated considering ten closest neighbors, resulting in giving groups of three accounts a high local outlier factor. Furthermore, it should be noted that both outlier detection methods, local outlier factor and cluster distance based score gives similar results over the whole group of accounts. Most accounts with a high score in one method

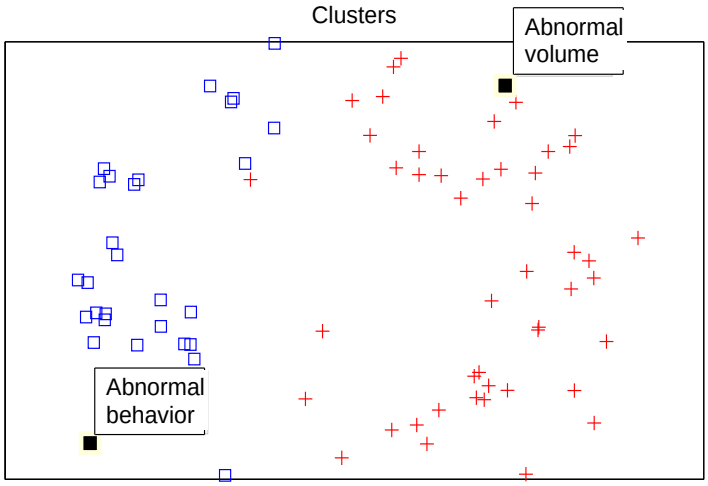

Figure 4.6: Clusters presented in two dimensions. Projection to two dimensions is done with multidimensional scaling. The distance criterion used in matlab clustering was *metrics stress*([32]).

also have a high score in the other.

To compare the efficiency of these two outlier detection methods, accounts were scored also by calculating the average distance from ten of their closest peers and this distance is presented as an outlier score (section 2.4.2). Figure 4.9 shows these scores. This method gives a high score for many accounts and fails to emphasize the importance of outlying behaviors as efficiently as clustering or local outlier factor methods. Although the created outliers receive a high score, the scores of several other accounts are close to the outliers.

The performance measure of the methods was discussed in section 3.5. Rank of the created outliers was stated as the performance measure in this context. Table 4.3 summarizes the performance of the different methods. The behavioral outlier was scored as the most suspicious account by all the methods. The volume based outlier was best detected with local outlier factors.

Table 4.3: Ranks of the created outliers in outlier analysis

Outlier type	**Clustering**	**LOF**	**PG distance**
Volume outlier	15	9	17
Behavioral outlier	1	1	1

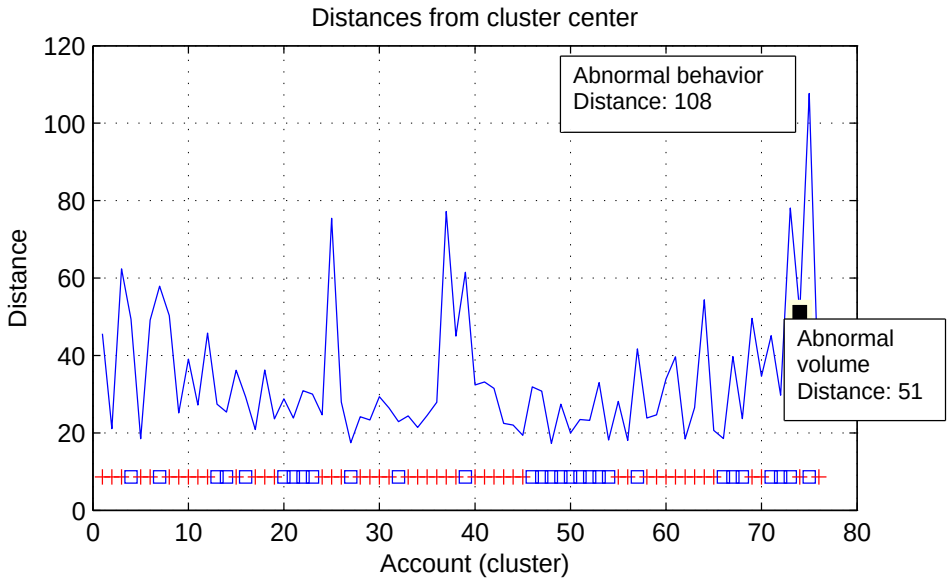

Figure 4.7: The outlier scores based on the distances from account cluster centers.

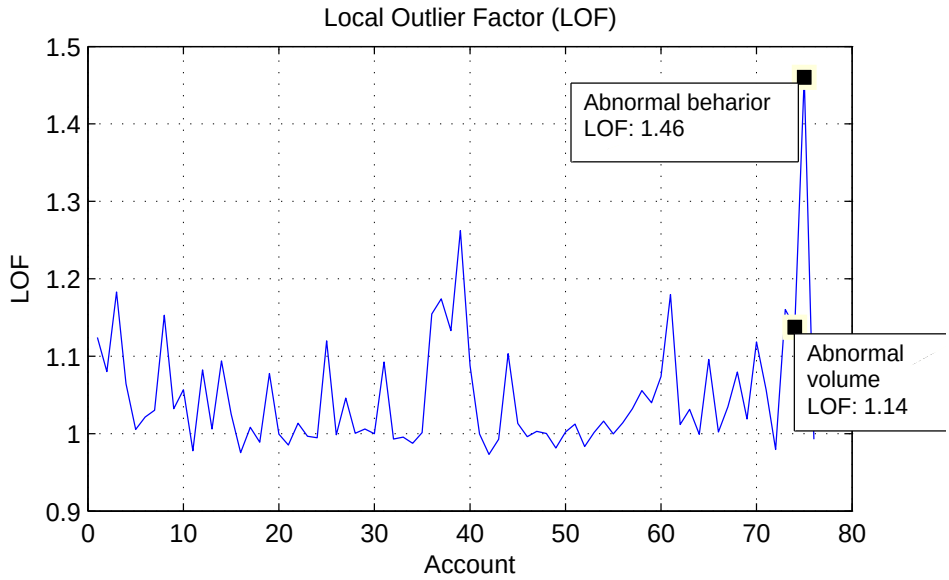

Figure 4.8: The outlier scores based on the local outlier factor. Ten closest neighbors are included in the calculation.

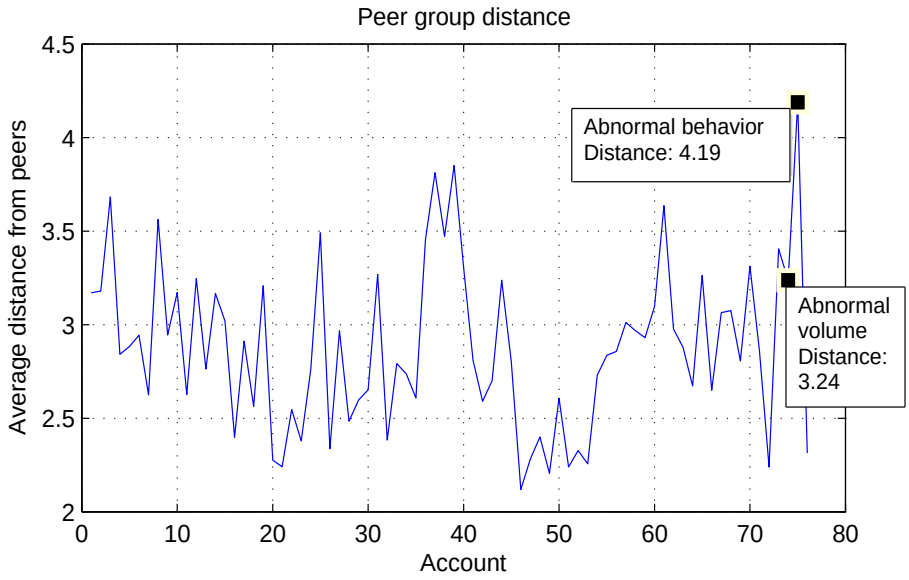

Figure 4.9: The outlier scores based on average distances from the closest peers. 10 closest peers are taken into account.

4.3 Behavioral change detection

After the initial outliers are discovered, the accounts that address a change in their behavior can be detected. The dataset is updated with 45 days of real testing data and the peer group analysis (sections 2.4.4 and 3.4) is run after each updated day to follow the suspiciousness of account changes. The peer group calculations are based on the total distance of accounts (3.3.3) and no weighting (2.3.1) was used in the update process.

The performance of behavioral change detection method is analyzed similarly as for the clustering and local outlier factor methods. Test accounts were created to represent different outlier types. There are four test accounts; one for sudden high volume in an attribute that has had only low volumes to represent an account with a single volume outlier, one for a repair type that the account has not performed earlier to represent an account that tries to copy the behavior of another account, one for continuous increase in repair volumes and one for change in distribution of repairs. More detailed description of creation of the outliers can be seen in Appendix A.2.

As discussed in section 3.4, the emphasis in the analysis is on detecting behavioral changes and thus the situation in the beginning of the analysis is not

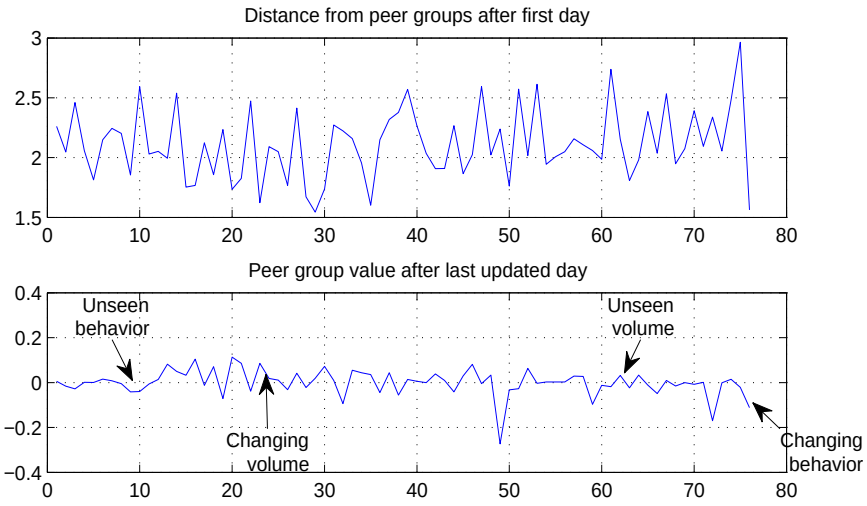

Figure 4.10: Account changes during the updating process. The arrows point out the peer group analysis values for each created suspicious account

relevant. Figure 4.10 shows both the starting situation and the level of change in the peer group analysis value. The upper graph shows how suspicious the accounts have been in the beginning of the analysis. The graph shows results that are very similar to the peer distance figure 4.9. The bottom graph shows the final peer group values from the analysis, calculated with (3.4.3). How this peer group analysis value evolved can be seen in figure B.2. Only the graphs for the five accounts that changed the most are shown. From the bottom graph in figure 4.10 it can be seen that account number 49 has changed the most compared to others. Even if the change in peer group value has been negative, meaning that the behavior of the account has moved closer towards its peers, the drivers behind this change should be investigated. Moreover, the graph shows that the created account with changing claiming behavior receives a high peer group analysis value. This account shows up as the fourth most suspicious account.

Figure 4.11 shows the changes of each signature component over time for the account number 49. The changes in each component of the signature are shown separately. The change of this account in two components is higher than for any other account. However, majority of this change has occurred only in three days of updates. These high changes imply that the historical volume of the account is low, and one day can impact the signature greatly. An examination

of the changes of the other accounts, either strongly dashed (peer accounts for account 49) or thin dashed accounts, shows that there are several other accounts that perform similarly, i.e., are having a great change during one day. Further analysis confirmed that when claims are updated to the account with only few historical data points, the signature is likely to change a lot.

The changes for the created outliers were also analyzed. These changes for the signature components of each outlier account are presented in figures B.4 - B.6 in Appendix B. The changes for these accounts stay moderate because all the accounts chosen for outliers had a proper historical signature before updating. Only the account with changing behavior the changes are significant but similar jumps as for account 49 cannot be seen. The account with increasing claiming volume starts to get increasing change values in components 2 and 3 as well as in the basic signature component. If the peer group analysis were to be continued longer, this account would appear in the results. This means that the volumes start to change significantly enough only after about 15 days of growth (see table A.1 for explanation of the outlier creation). Because the level of change appears to be insufficient, the outlying account could be regenerated with a greater change. However, in real world fraud cases, account volumes do not have drastic changes, but the volumes increase slowly. Thus great volume increase would not be justified in the analysis.

The accounts that have single suspicious transactions are not identified as outliers. The account with sudden high volume (figure B.6) had the increase in volume on day 25. This appears as a change only in component 1 and component 2, although similar changes can be observed already within the original transactions of the account. The account with claims that it has never claimed earlier does not show significant changes in component values either (figure B.7).

Table 4.4: Simulation results for the created outliers

Outlier type	**PG distance**	**PG analysis**
Changing volume	(36)	47
Changing behavior	(75)	4
Volume outlier	(30)	35
Behavior outlier	(4)	29

Table 4.4 summarizes the peer group analysis performance. The rank of each

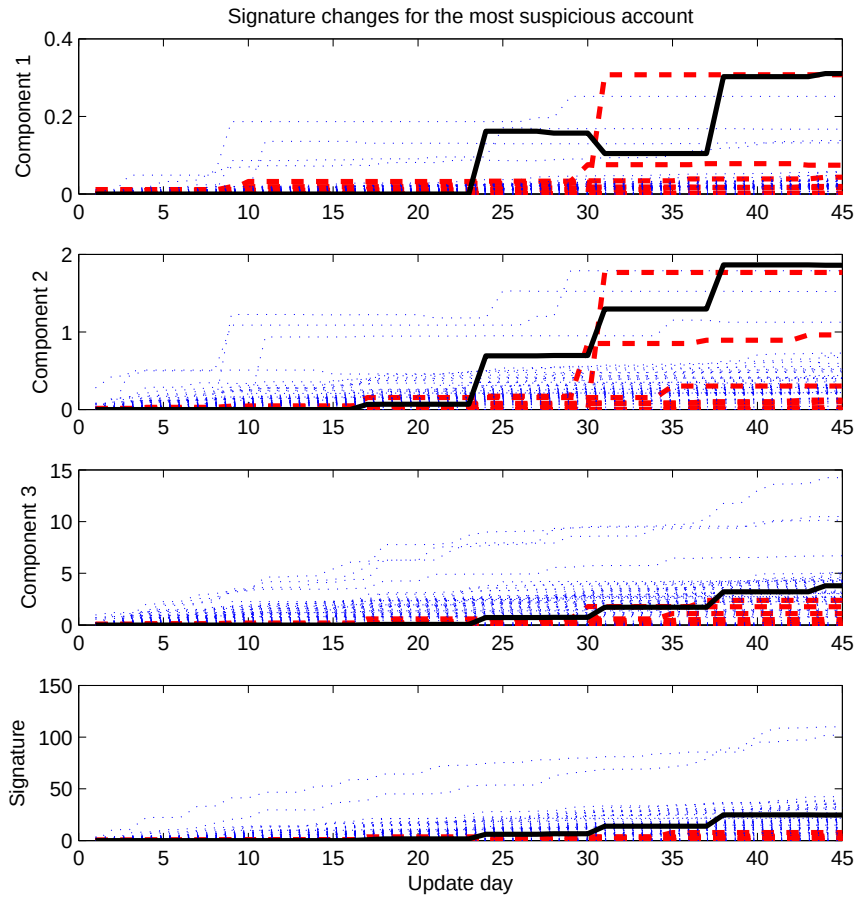

Figure 4.11: Account changes for the account that changed the most according to peer group analysis. Changes are presented for each signature component separately. The strongly dashed lines present the changes of the peer accounts and the shallow dashed lines present all the other accounts.

tested outlier account is presented in the second column. The peer group distance ranks are presented as a reference. The table emphasizes the fact that peer group analysis fails to identify other outliers except the behavioral change outlier. This outlier is detected well and is pointed out as the fourth suspicious account.

Chapter 5

Summary and conclusions

This thesis presented a fraud detection method for warranty claim fraud. The presented method was tailored to fit the existing capabilities and needs of an international electronics manufacturing company. The aim for the method was to detect those repair vendors whose repairing behavior is more suspicious than others'. Identifying the vendors with possible fraudulent behavior would bring the company monetary benefits in terms of prevented losses.

The approach taken in this thesis was based on the assumption that, from statistical point of view, when no special rules apply, no repair partner is unique by performance. Based on this, histogram based signatures that profile each partner separately, were presented. Histogram signatures are histogram probability distributions of selected dimensions. Histogram signatures fit well to the data and bring a new approach to looking at the data with conditional histograms of daily repair volumes. The advantages of this method are simplicity and efficient use of memory, while the only thing that needs to be saved is the histogram bin value. However, already with the data used in this thesis the number of possible conditioning categories was so high that the computations were quite laborious. If the number of categories increases much, the similarly behaving partners should be divided into partial analysis groups. This, on the other hand, would decrease the visibility that including all partners to one analysis brings.

The selection of categories has been emphasized not only relating to computational efficiency but also in relation to the reliability of the results. The χ^2 test that was presented with the variable selection is not considering all

the accounts and categories when making the decision of including a variable. The empty histograms and empty mutual bins of histograms are removed in the beginning of the test for each account [12]. The data used in the thesis is very scattered, and for many accounts several categories are empty. If only for example ten categories out of 1535 categories create a positive result, the result is not very valid and this type of account will receive a high importance in change detection because every new observation may shift the histogram distribution greatly. Another type of test would be needed that would take also the change possibility into account.

The methods presented in this thesis only succeed to identify certain type of outliers. It detects deviation based outliers well, meaning the accounts for which the behavior is different from others. This applies to both phases; to the initial analysis where the accounts were analyzed with clustering and nearest neighbor methods, and to the peer group analysis where the change in behavior was calculated against the accounts own behavior, i.e. signature. However, the results imply that the deviation from others must be wide. If the deviation only appears in one specific account characteristics, normal behavior in other characteristics rules out the strange behavior. The same implies also for the single outliers, whose impact is not sufficient to raise suspicion for longer period of time.

Histogram signatures tend to identify abnormalities and changes well when the number of variable attributes is low. In this case, the differences can be seen easily. For instance, although not written in the results section, during the thesis signature analysis on one variable proved to be more valuable than analysis where all the variables were included. This thesis did not concentrate on only one signature component, because the total signature differences were analyzed, but the emphasis should be put to one component at a time in the future.

Although the methods could not identify all the created outliers, it should be kept in mind that the creation of outliers could have been flawed also. For instance, the formulation could have been more random. Now the creation was as random as possible, but all the outlying accounts that needed to be created using the original set of data were invented, resulting in possibly unrealistic outliers. On the other hand, as the used data was actual data from a manufacturer's database, all the accounts themselves can already include outlying

behavior.

An interesting topic for future development of the method is how to identify the most important changes from the manufacturing company's perspective. For example some type of weighting system for the most expensive changes or attributes is worth studying. This thesis presented only one way to weight the different components of the signature, by improving the weighting the performance of the method may improve a lot.

From financial perspective an important improvement would be to include the cost savings dimension to the analysis. It was seen in the chapter 4 that the peer group analysis now identifies best the changes in small volume accounts. Very often these accounts are not the most interesting. For cost optimization, some type of price tag could be included to the peer group analysis, by analyzing the cost of daily repairs instead of repair volumes for instance.

To conclude, this thesis succeeded to fulfill its objectives. A new approach to warranty fraud detection was presented and the method is able to detect the most important fraud indicators. The method has already been put in practice in the manufacturing company and it has managed to identify some major differences in repair vendors.

Bibliography

[1] Shin Ando. Clustering needles in a haystack: An information theoretic analysis of minority and outlier detection. *IEEE International Conference on Data Mining*, pages 13–22, 2007.

[2] Andreas Arning, Rakesh Agrawal, and Prabhakar Raghavan. A linear method for deviation detection in large databases. In *Knowledge Discovery and Data Mining*, pages 164–169, 1996.

[3] Pavel Berkhin. Survey of clustering data mining techniques. Technical report, Accrue Software, Inc.

[4] Richard J. Bolton and David J. Hand. Peer group analysis - local anomaly detection in longitudinal data. Technical report, Department of Mathematics, Imperial College, London, 2001.

[5] Richard J. Bolton and David J. Hand. Unsupervised Profiling Methods for Fraud Detection. *Statistical Science*, 17(3):235–255, 2002.

[6] Shyam Boriah, Varun Chandola, and Vipin Kumar. Similarity measures for categorical data: A comparative evaluation. In *SIAM International Conference on Data Mining*, pages 243–254, 2008.

[7] Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng, and Jörg Sander. Lof: Identifying density-based local outliers. In *International Conference on Management of Data*, pages 93–104, 2000.

[8] Michael Cahill, Diane Lambert, José C. Pinheiro, and Don Sun. Detecting fraud in the real world. In *Handbook of massive data sets*, pages 911 – 929. Kluwer Academic Publishers, Norwell, MA, 2000.

[9] Huseyin Cavusoglu, Birendra Mishra, and Srinivasan Raghunathan. The value of intrusion detection systems in information technology security architecture. *Information Systems Research*, 16(1):28–46, March 2005.

[10] Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM Computing Surveys*, 41:1–58, 2009.

[11] Varun Chandola, Shyam Boriah, and Vipin Kumar. Understanding categorical similarity measures for outlier detection. Technical report, Computer Science Department, University of Minnesota, 2008.

[12] Fei Chen, Diane Lambert, Carla Pinheiro, and Don X. Sun. Reducing transaction databases, without lagging behind the data or losing information. *Bell Labs Technical memorandum*, 2000.

[13] Fei Chen, Diane Lambert, and José C. Pinheiro. Incremental quantile estimation for massive tracking. In *Proceedings of KDD*, pages 516–522, 2000.

[14] Tom Fawcett and Foster Provost. Adaptive fraud detection. *Data Mining and Knowledge Discovery*, 1:291–316, January 1997.

[15] Tom Fawcett and Foster J. Provost. Activity monitoring: noticing interesting changes in behavior. In *Knowledge Discovery and Data Mining*, pages 53–62, 1999.

[16] Fraud. `http://dictionary.reference.com/browse/fraud`, visited 19.11.2011.

[17] Ana L. N. Fred and Anil K. Jain. Robust data clustering. In *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 128–133, 2003.

[18] David Froning. Detecting claim fraud and improving product quality: Case studies in reducing warranty costs. In *SAS Conference Proceedings: SAS Global Forum 2010*. SAS Institute Inc., Cary. NC, 2010.

[19] Brian R. Gaines. System identification, approximation and complexity. *International Journal of General Systems*, 3:145–174, 1977.

[20] Jim Gee. Fraud 2009 bad, 2010 better? *Computer Fraud & Security*, February 2010.

[21] Sigi Goode and David Lacey. Detecting complex account fraud in the enterprise: The role of technical and non-technical controls. *Decision Support Systems*, 50(4):702–714, 2011.

[22] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer New York Inc., New York, NY, USA, 2001.

[23] Douglas M. Hawkins. *Identification of Outliers*. Chapman and Hall, London, 1980.

[24] Zengyou He, Xiaofei Xu, and Shengchun Deng. Discovering cluster-based local outliers. *Pattern Recognition Letter*, 24:1641–1650, June 2003.

[25] Zengyou He, Xiaofei Xu, and Shengchun Deng. A fast greedy algorithm for outlier mining. *The Computing Research Repository (CoRR)*, abs/cs/0507065, 2005.

[26] Zengyou He, Xiaofei Xu, and Shengchun Deng. An optimization model for outlier detection in categorical data. *The Computing Research Repository (CoRR)*, abs/cs/0503081, 2005.

[27] Edwin M. Knorr and Raymond T. Ng. Algorithms for mining distance-based outliers in large datasets. In *Proceedings of the 24th International Conference on Very Large Data Bases*, pages 392–403, New York, NY, 1998.

[28] Edwin M. Knorr and Raymond T. Ng. Finding intensional knowledge of distance-based outliers. In *Proceedings of the 25th International Conference on Very Large Data Bases*, VLDB '99, pages 211–222, San Francisco, CA, USA, 1999. Morgan Kaufmann Publishers Inc.

[29] Pertti Laininen. *Todenäköisyys ja sen Tilastollinen Soveltaminen*. Oy Yliopistokustannus / Otatieto, Helsinki, 2004.

[30] Diane Lambert, José C. Pinheiro, and Don X. Sun. Updating timing profiles for millions of customers in real-time. *Bell Labs Technical mem-*

orandum, 1999.

[31] Dekang Lin. An information-theoretic definition of similarity. In *Proceedings of the 15th International Conference on Machine Learning*, pages 296–304. Morgan Kaufmann, 1998.

[32] MATLAB. `http://www.mathworks.se/help/techdoc/ref/f16-6011.html`, visited 19.07.2012 (require a registration).

[33] Harvey Motulsky. *Intuitive Biostatistics: Choosing a Statistical Test*. Oxford University Press, 1995.

[34] Kevin P. Murphy. Elements of information theory. In *Advances in Knowledge Discovery and Data Mining, Fayyad, U.* Wiley, 1998.

[35] Frank Nielsen. A family of statistical symmetric divergences based on jensen's inequality. *The Computing Research Repository (CoRR)*, abs/1009.4004, 2010.

[36] Matthew Eric Otey, Amol Ghoting, and Srinivasan Parthasarathy. Fast distributed outlier detection in mixed-attribute data sets. *Data Mining and Knowledge Discovery*, 12:2–3, 2006.

[37] Qlikview. `http://www.qlikview.com/fi/explore/products/overview`, visited 07.05.2012.

[38] MATLAB R2011b. `http://www.mathworks.se/products/matlab`, visited 30.06.2012.

[39] Gregory E. Sims, Se-Ran R. Jun, Guohong A. Wu, and Sung-Hou H. Kim. Alignment-free genome comparison with feature frequency profiles (ffp) and optimal resolutions. *Proceedings of the National Academy of Sciences of the United States of America*, 106(8):2677–2682, February 2009.

[40] Inder Jeet Taneja. On generalized information and divergence measures and their applications: A brief review. *Qüestiió*, 13(1-3):47–73, 1989.

[41] Inder Jeet Taneja. *Generalized Information Measures and Their Applications*. `http://www.mtm.ufsc.br/~taneja/book/book.html`, 2001.

[42] Fraud types. `http://www.lookstoogoodtobetrue.com/fraud.aspx`,

visited 19.11.2011.

[43] Yuan Wei. Variance, entropy and uncertainty measure. *American Statistical Association 1987 Proceedings(with Liu Leping)*, 1987. http://www.amstat.org/sections/SRMS/proceedings/papers/1987_108.pdf.

[44] David J Weston, David J Hand, Niall M Adams, Christopher Whitrow, and P Juszczak. Plastic card fraud detection using peer group analysis. *Advances in Data Analysis and Classification*, 2(1):45–62, 2008.

Appendix A

Explanations

A.1 Definitions

Distribution based outliers. Data is assumed to follow some standard distribution, and the outlier factor for each data instance is calculated against this distribution. These methods are most useful when the data collection process is strictly controlled [26].

Distance based outliers. Objects that are distant from other data [27]. The number of compared data objects varies, some techniques calculate the distance to a limited number of closest neighbors, while other techniques take the distance to all data objects into account.

Deviation based outliers are found by studying the characteristics of data objects and identifying the objects with deviating features [26].

Density based outliers. Local outlier factor (LOF) that points data objects with scarce neighborhoods as outliers was first presented by Breunig et al. [7].

Depth based outliers appear when data objects are layered in the data space and the layers are compared against each other. Outliers are assumed to lie in shallow layers [28].

Clustering based outliers can be identified with some clustering method that is able to detect small and large clusters. Objecst in small clusters or clusters relatively far from others are regarded as outliers [24].

Pearson's chi-squared (χ^2) test can be used to compare the statistical significance of the differences in datasets. Traditionally χ^2 test studies if a dataset is following some chosen hypothesis [29].

A.2 Created anomalous accounts

Table A.1: Test case creation for simulations. The performance of the parent accounts used in the the creation of the test accounts can be compared against the performance of the created accounts. The method should rank the created accounts with a smaller rank than the parent accounts.

ISSUE	**Volume related issue**	**Behavior related issue**
Abnormality	The data for account 1 is increased with the datapoints that have fairly high volume. This new account is marked as account 74.	Combination of accounts that have reported only some of the first seven attributes for variable 2. This combined account was given the account number 75.
Change related	The repair activities for account 24 are increased slightly from their normal level on each day it performs repairs. The increasing begins after day 10 and each day the repairs are increased with 10%.	Accounts 11 and 26 are combined before peer group analysis and made to perform like account 28 after the behavior analysis starts. This combination of accounts has the account number 76.
Single outlier	The account for this outlier was chosen randomly. Account 62 was chosen. For a random day, chosen day 25, the maximum volume of that account was increased 10 times.	The account was chosen randomly. Account 10 was chosen. For a random day, chosen day 19, the account was given transactions it had not reported earlier.

Appendix B

Figures

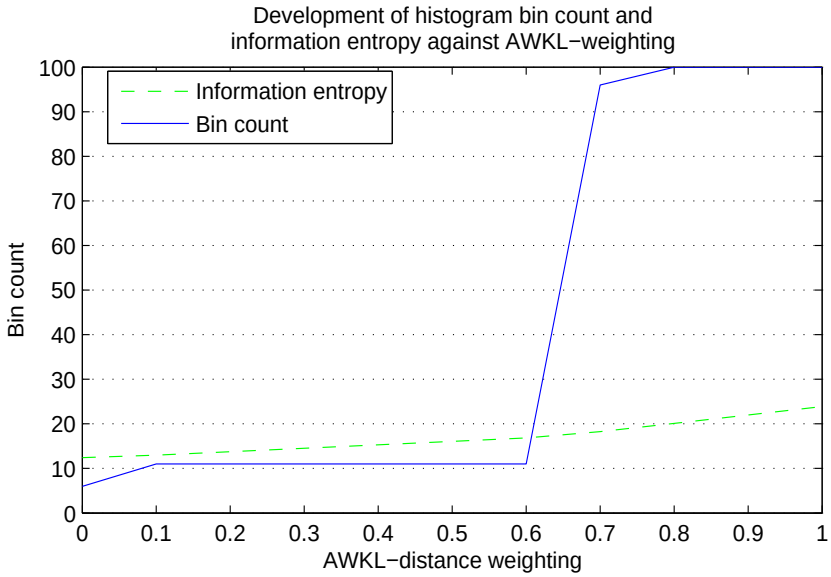

Figure B.1: Development of the bin count and information entropy against weighting parameter w.

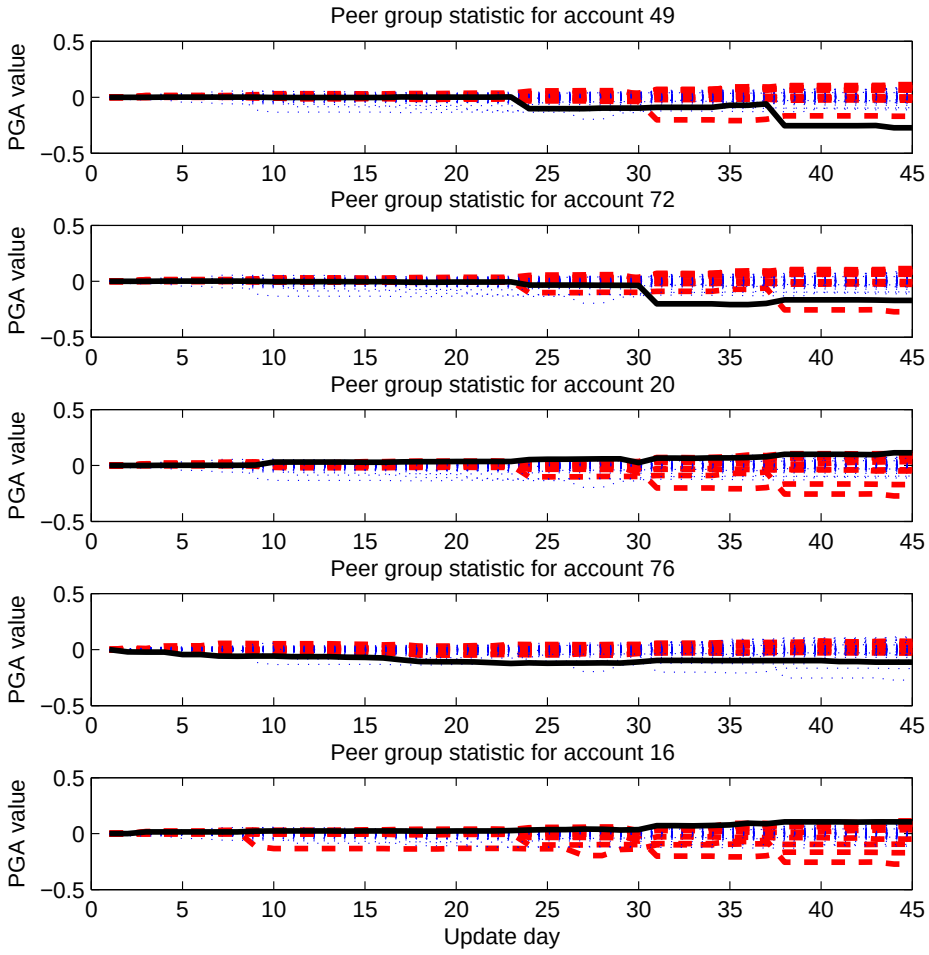

Figure B.2: Peer group analysis (PGA) results for the five most changed accounts.

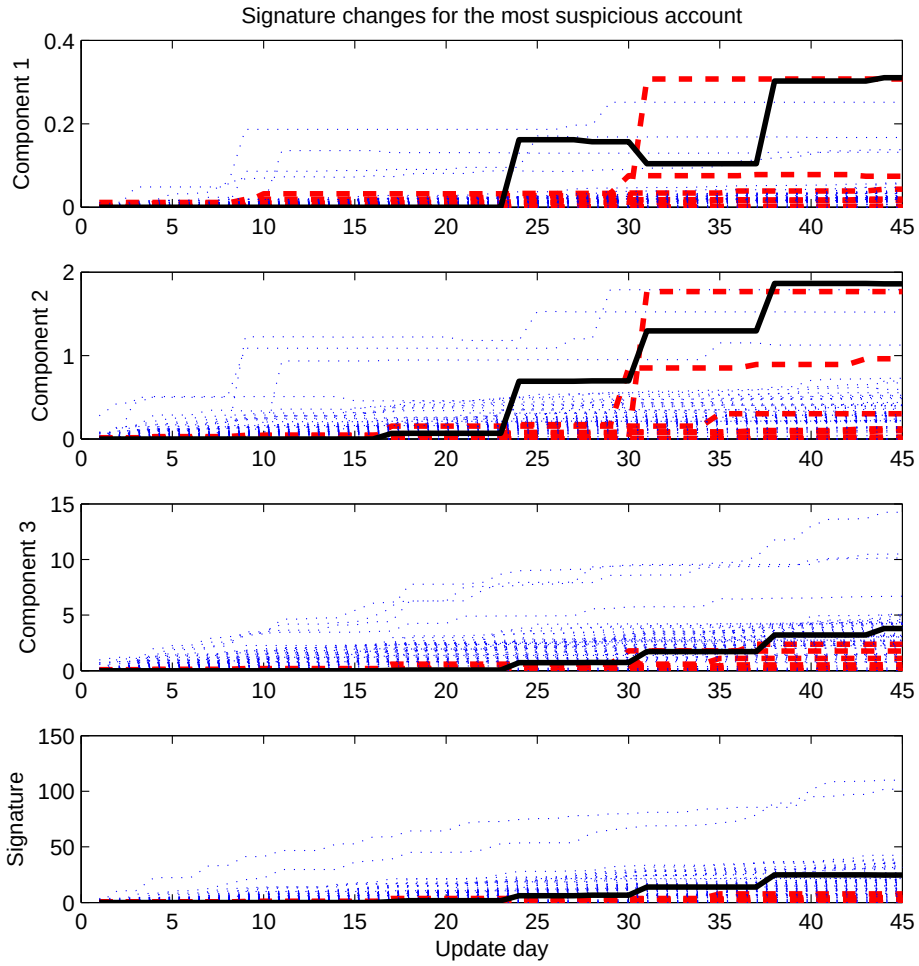

Figure B.3: Account changes for the account that changed the most according to peer group analysis(account 49). Changes are presented for each signature component separately. The strongly dashed lines present the changes of the peer accounts and the shallow dashed lines present all the other accounts.

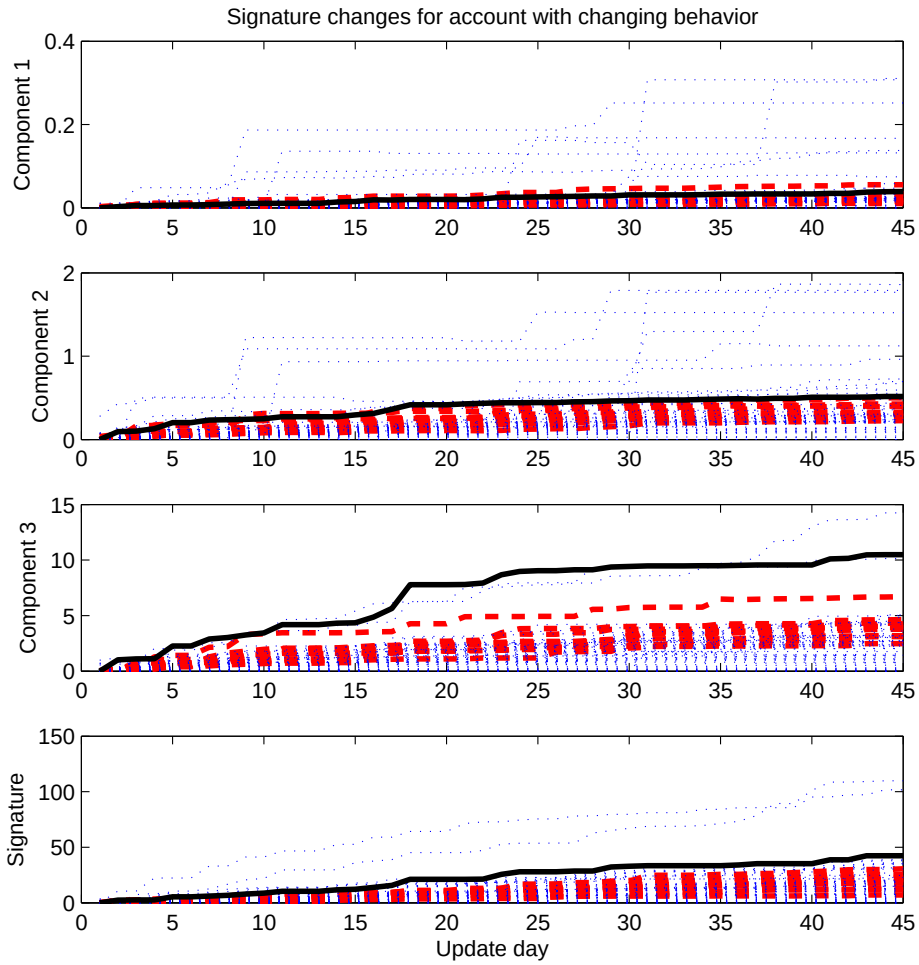

Figure B.4: Account changes for the created outlier that changes the behavior right from the beginning of the peer group analysis (account 76). Changes are presented for each signature component separately. The strongly dashed lines present the changes of the peer accounts and the shallow dashed lines present all the other accounts.

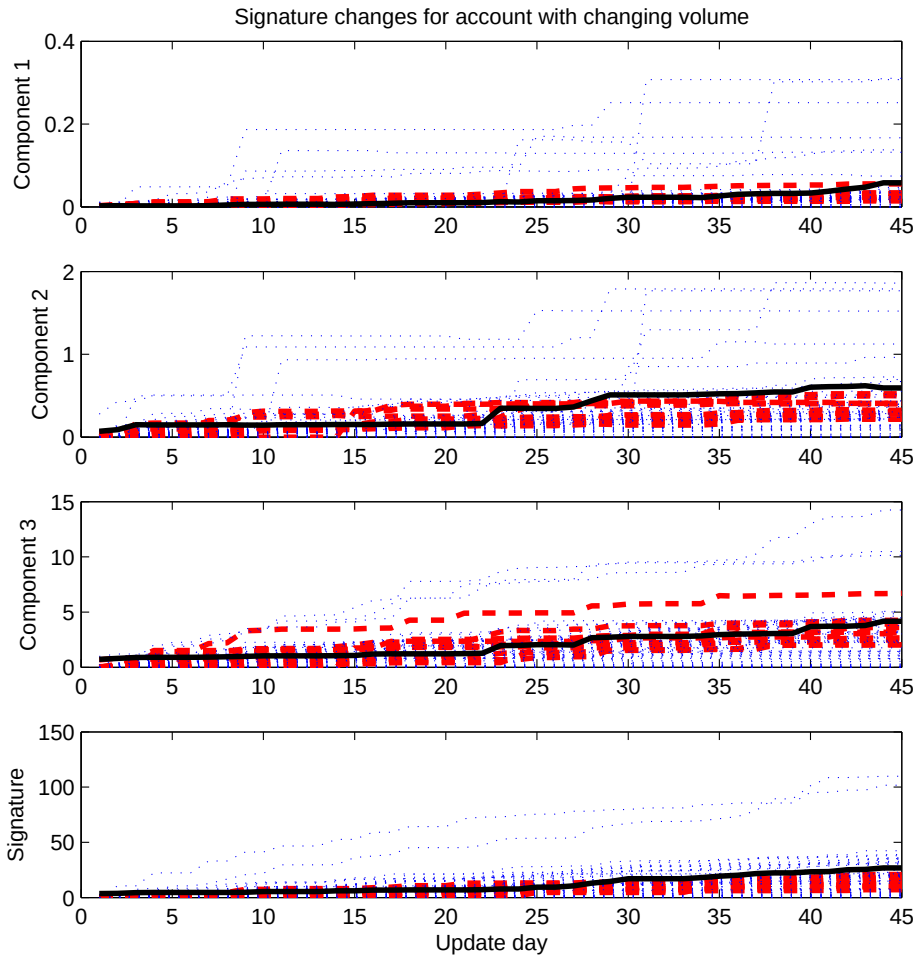

Figure B.5: Account changes for the created outlier with changing volume (account 24). Changes are presented for each signature component separately. The strongly dashed lines present the changes of the peer accounts and the shallow dashed lines present all the other accounts.

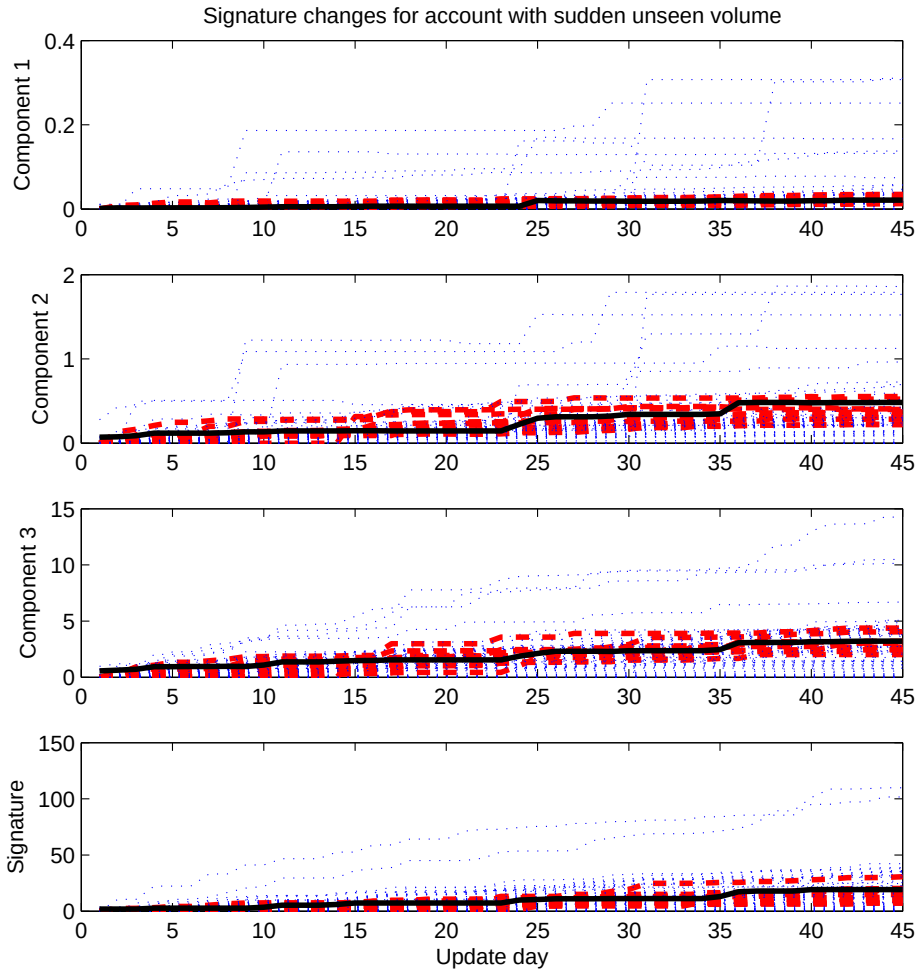

Figure B.6: Account changes for the account that had a single day with surprisingly high claim volume (account 62). Changes are presented for each signature component separately. The strongly dashed lines present the changes of the peer accounts and the shallow dashed lines present all the other accounts.

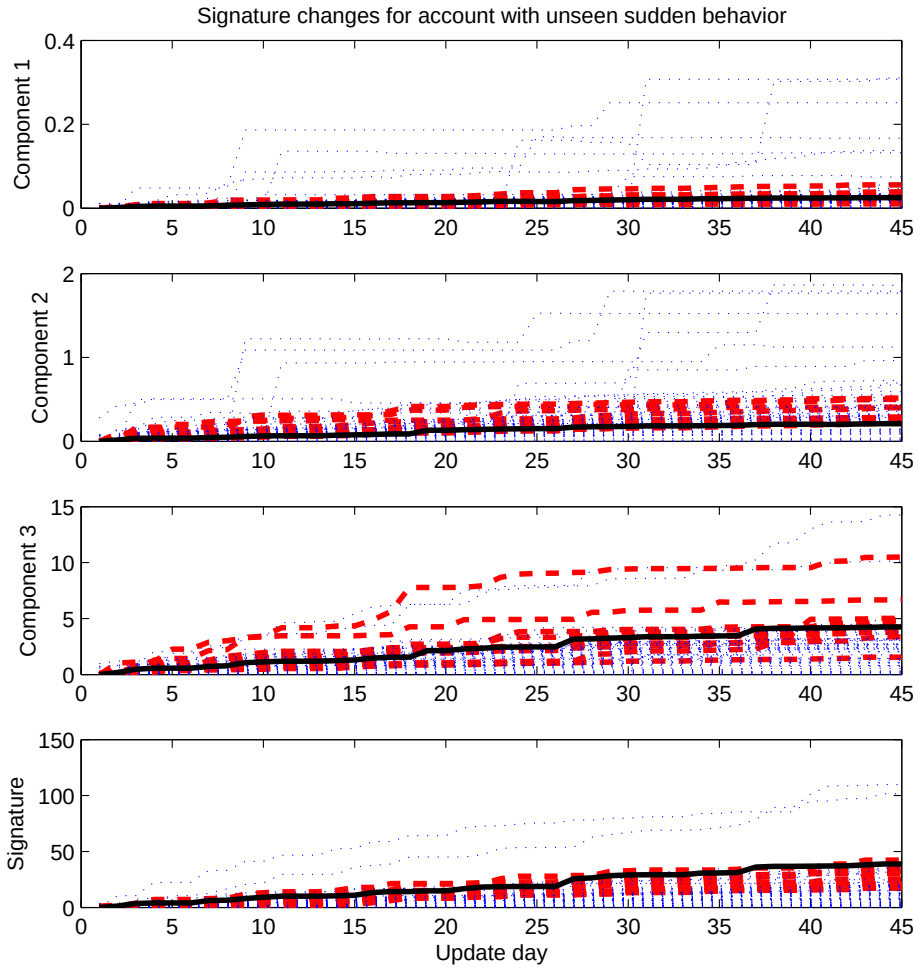

Figure B.7: Account changes for the account that had a single day with surprising claims (account 10). Changes are presented for each signature component separately. The strongly dashed lines present the changes of the peer accounts and the shallow dashed lines present all the other accounts.